State of the art in speech and audio processing

Khalid Daoudi

September 20, 2004

Chapter 1

State of the art in Acoustic Speech Recognition

K. Daoudi and D. Fohr
INIRIA-Parole, France

1.1 Statistical speech recognition

The goal of an automatic speech recognition system is to deduce meaningful linguistic units (i.e., words) from acoustic waveforms. Due to the random nature of the process and interferences, it is not possible to derive a deterministic formulation that provides a mapping between acoustic signal and conceptual meanings. Instead the problem is generally formulated in a probabilistic framework. In probabilistic setting the speech recognition is stated as the estimation of 'most probable' linguistic representation of a 'given' acoustic waveform. The mathematical formulation of this problem is:

$$\hat{W} = \arg\max_{W} P(W|O)$$

where O is a set of observations from the acoustic waveform and W is a random variable that takes its values from the possible linguistic representations in the language under consideration. $P(W|O)$ is the conditional probability distribution of the linguistic representation given the observations. This conditional distribution constitutes the knowledge base of the recognizer. This knowledge base is constructed using statistical learning techniques and a priori expertise on speech production mechanisms. The role of a priori expertise on the domain is to provide a set of simplification assumptions that will guide the statistical machinery to extract relevant information for recognition.

Because of the complexity of the speech production mechanisms there is no simple parametric representation of $P(W|O)$ that involves both acoustic and linguistic information. The basic approach is to first divide the problem into acoustic and linguistic components that can be handled separately. This is achieved using a Bayesian reformulation:

$$\hat{W} = \arg\max_{W} P(O|W)P(W).$$

In this formulation, the acoustic model, $P(O|W)$, encodes the statistical distribution of speech acoustics given the linguistic labeling. $P(W)$ is the probability assigned by the language model which encodes the a priori linguistic information. This approach is called the source-channel model. $P(W)$ constitute the language source with all the linguistic constraints of the underlying language and $P(O|W)$ is the acoustic channel that outputs the speech signal based on the linguistic unit W. With this formulation the two components of the system can be constructed separately later to be combined in the recognition phase. The information source for the linguistic component is written text documents that is supposed to be sufficient to represent the properties of language. For the acoustic component it is the speech utterances

labeled with corresponding linguistic units, i.e. words, phonemes. For both of these components the statistical approach consists of extraction of relevant information content from data and formulation of a parametric representation that is capable to encode this content.

1.2 Acoustic Modeling

State-of-the-art automatic speech recognition (ASR) systems are based on probabilistic modeling of the speech signal using Hidden Markov Models (HMM). HMM are probabistic automaton. The goal of decoding process is to determine a sequence of states that the observed signal has gone through. There are three main problems:

- the Evaluation problem: given a model and a sequence of observations, what is the probability what the model generated the observations? [9]
- the Decoding problem: given a model and a sequence of observations, what is the most likely states sequence that produced the observation ? [11]
- the Learning problem: given a model and a sequence of observations, what should the model's parameters be so it has the maximum probability of generating the observations ? [7]

Recently, the acoustic modeling problem in speech recognition was reformulated within the probabilistic graphical models (PGM) formalism [3]. PGM is a unifying framework for statistical learning which provides an abstraction of quantitative and qualitative components of a statistical model. Dynamic Bayesian networks (DBN) are a subset of PGM which are defined on directed acyclic graph structures and which include HMM as a special case [4]. These models are defined with graph structures that encode the probabilistic relations between its variables through a set of associated conditional probabilities. One of the main advantages of PGM is the graphical abstraction that provides a visual understanding of the modeled process. Moreover they provide a powerful setting to specify efficient inference algorithms that can be specified automatically once the initial structure of the graph is determined.

1.3 Speech feature extraction

In feature extraction stage, the speech signal is considered as a quasi-stationary process consisting of consecutive frames that can be treated independently.

The goal of front-end speech processing in ASR is to attain a projection of the speech signal to a compact parameter space where the information related to speech content can be extracted easily. Most parameterization schemes are developed based on the source-filter model of speech production mechanism [9, 10]. In this model, speech signal is considered as the output of a filter (vocal tract) whose input source is either glottal air pulses or random noise. For voiced sounds the glottal excitation is considered as a slowly varying periodic signal. This signal can be considered as the output of a glottal pulse filter feed with a periodic impulse train. For unvoiced sounds the excitation signal is considered as random noise.

State of the art speech feature extraction schemes (Mel frequecy cepstral coefficients (MFCC) and perceptual linear prediction (PLP)) are based on auditory processing on the spectrum of speech signal and cepstral representation of the resulting features [2]. The spectral and cepstral analysis is generally performed using Fourier transform. The advantage of Fourier transform is that it possesses very good frequency localization properties.

1.3.1 Linear Prediction Coefficients

Linear predictive coding (LPC) has been considered one of the most powerful techniques for speech analysis. LPC relies on the lossless tube model of the vocal tract. The lossless tube model approximates

the instantaneous physiological shape of the the vocal tract as a concatenation of small cylindrical tubes. The model can be represented with an all pole (IIR) filter. LPC coefficients a_k can be estimated using autocorrelation or covariance methods [10].

1.3.2 MFCCs (Mel-Frequency Cepstral Coefficients)

Cepstral analysis denote the unusual treatment of frequency domain data as it were time domain data [1]. The cepstrum is a measure of the periodicity of a frequency response plot. The unit measure in cepstral domain is second but it indicates the variations in the frequency spectrum.

One of the powerful properties of cepstrum is the fact that any periodicities, or repeated patterns, in a spectrum will be mapped to one or two specific components in the cepstrum. If a spectrum contains several harmonic series, they will be separated in a way similar to the way the spectrum separates repetitive time patterns in the waveform. The mel-frequency cepstral coefficients proposed by Mermelstein [2] make use of this property to separate the excitation and vocal tract frequency components in cepstral domain. The spectrum of excitation signal is composed of several peaks at the harmonics of the pitch frequency. This constitutes the quickly varying component of the speech spectrum. On the other hand the vocal tract frequency response constitutes the slowly varying component of the speech spectrum. Hence a simple low pass liftering (i.e. filtering in cepstral domain) operation eliminates the excitation component.

1.4 Noise compensation

In ASR systems one of the commonly encountered problems is the mismatch between training and application conditions. Solutions to this problem are provided as pre-processing of the speech signal for enhancement, noise resistant feature extraction schemes and statistical adaptation of models to accommodate application conditions. In order to achieve improved performances all these techniques should be applied in a practical system. In compensation schemes the matching between training and application conditions is achieved either using a hypothesized mismatch function that describes the noise corruption or applying statistical adaptation of the model parameters using data recorded in new environment. In adaptive compensation such as MAP [6] and MLLR [8], the observed corrupted speech data is used to transform the initial models. One inconvenience of these schemes is that they require at least few dozens of seconds of speech to yield relatively good estimates of the new models. A more important inconvenience is the need of speech data transcription in an unsupervised mode, this is done generally using the initial models which limits the performances if the transcription accuracy is poor. Predictive compensation schemes (PMC [5] for instance) do not rely on observed speech data, rather they use noise observations and models to estimate the speech models in the new environment. Namely, the new speech model is a combination of the initial one and a (parametric) noise model for which the parameters are estimated from noise observations. The combination is obtained using a function (mismatch function) that hypothesizes the corruption of speech by noise sources [5]. An inconvenience of predictive schemes is the assumption made about the effects of the new acoustic environment on the clean speech, i.e., how the noise sources alter the clean speech signal (additive or/and convolutional...). This assumption may be unrealistic which consequently may lead to erroneous estimations of corrupted speech distributions. An other inconvenience of predictive schemes is their strong dependence on the front-end (MFCC in general) and the probabilistic models (HMMs in general) which are used.

Bibliography

[1] B.P. Bogert, M.J. R. Healy, and J.W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, psuedo-autocovariance, cross-cepstrum and sa phe cracking. In *Proceedings of the Symposium on Time Series Analysis*, 1963.

[2] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken senten ces. *IEEE Transactions on Acoustics, Speech and Signal Processing*, (4):357–366, 1980.

[3] J. Bilmes. Graphical Models and Automatic Speech Recognition. In *Mathematical Foundations of Speech and Language Processing*, Institute of Mathematical Analysis Volumes in Mathematics Series, Springer-Verlag, 2003.

[4] K. Daoudi and D. Fohr and C. Antoine. Dynamic bayesian networks for multi-band automatic speech recognition. *Computer Speech and Language*, Vol 17, 2003. pp.263-285.

[5] M.J.F Gales. Predictive model-based compensation schemes for robust speech recognition. *Speech Communication*, 25(1-3):49–74, 1998.

[6] J.L. Gauvain and C.H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech and Audio Processing*, 2:291–298, 1996.

[7] R. Mercer L. Bahl, F. Jelinek. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 5(2):179–190, 1983.

[8] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hmms. *Computer Speech and Language*, 9:171–186, 1995.

[9] F. Rabiner and B.H. Juang. *Fundamentals of speech recognition*. Pearson Education POD, 1993.

[10] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, Englewood Cliffs, NJ, USA, 1978.

[11] A. Viterbi. Error bounds for convolutional codes and asymptotically optimum decoding algorithm. *IEEE Trans. Information Theory*, IT-13:260–269, 1967.

Chapter 2

Robust Features for Automatic Speech Recognition Systems

D. Dimitrios, P. Vassilis and P. Maragos
ICCS-NTUA, Greece

2.1 Introduction

Since the early scientific or fictitious envisagements of intelligent machines at Bell Labs, Lincoln Labs, or Clarke and Kubrick's Space Odyssey, computer systems have become ubiquitous, storing huge quantities of multimodal data (combinations of speech, audio, text and video). Managing this information is essential for the creation of a knowledge-driven society. Towards this direction, Automatic Speech Recognition (ASR) seems to be one of the most important tasks that should be successfully dealt with.

ASR technology has developed rapidly. The pioneering work of the first research years (Filterbanks, Spectrogram, Linear Prediction Coding) has been followed by several fundamental achievements (Dynamic Time Warping, Mel-Cepstral Coefficients, Hidden Markov Models). Although significant contributions have been made, ASR systems haven't yet reached the desirable standards of functionality in ordinary (everyday) conditions. Robustness is one of the main attributes that ASR systems lack. This could be tackled by the application of speech enhancement techniques, extraction of robust features for speech representation, or, finally, by model compensation.

As far as feature extraction is concerned, the main research areas cannot be easily classified in completely distinct categories, since the cross-fertilization of ideas has triggered approaches that combine ideas from various fields. *Filterbank analysis* is an inherent component of many techniques for robust feature extraction. It is inspired by the physiological processing of speech sounds in separate frequency bands that is performed by the auditory system. *Auditory processing* has developed into a separate research field and has been the origin of important ideas, related to physiologically and perceptually inspired features [16, 27, 51],[19, 21]. Equally important is the research field based on concepts relevant to speech resonance (short-term) *modulations*. Both physical observations and theoretical advances support the existence of modulations during speech production [52, 30],[37]. Other approaches are related to the long-term *modulation spectrum* [26, 11] and the features derived from that [23],[21], [22], which could be perceptually based or variants of noise robust features. Finally, special attention should be paid to the techniques [34, 38, 40, 4] that attempt to model nonlinear phenomena of the speech production system [52, 30]. These may quantify aerodynamic phenomena like turbulence and/or modulations, that the linear source-filter model cannot take into consideration.

2.2 Filterbanks

The corresponding group of ASR features is based on the idea of decomposing speech along the frequency domain using several overlapping bandpass filters. The filterbank scheme is motivated by observations made by Allen [2] and Fletcher [13]. They have provided evidences that the human auditory system processes speech in separate frequency bands and extracts their spectral content. The human cognitive system classifies the speech events accordingly. The most common features for ASR tasks are the time-localized energies of the different frequency bands, [39]. These features map the spectral subband energies to the appropriate acoustic events (phonemes). A common practice is to train separate recognizers to process each one of the band components.

For the definition of a filterbank certain parameters are required. These parameters are the number of filters, their placing (the center frequencies), their bandwidths and the type of filters used. The number of filters cannot be too small otherwise the ability to resolve the speech spectrum could be impaired. Also, their number cannot be too large because the filter bandwidth would be too small and some bands would have very low speech energy. The most common range for the number of filters is between 6 and 32 [48]. The filter placing can be linear where the center frequencies are spaced uniformly to span the whole frequency range of the speech signals. An alternative to uniform filterbanks is to space the filters uniformly along a logarithmic frequency scale (e.g. the Mel scale). Such a logarithmic frequency scale is motivated by the human auditory perception process. Finally, a common non-uniform filterbank placing is the critical band scale (Bark scale). The spacing of the filters along the critical band scale is based on perception studies and is intended to choose bands that give equal contribution to speech articulation [48]. The third parameter is the filters' bandwidths, which depend on the placing, the number and the desired overlap of the filters. It is common practice that the filter bandwidths are not equal along the frequency axis. Finally, many different types of bandpass filters have been proposed during the past few years depending on the analysis/recognition tasks. For instance, gammatone filters are popular for auditory speech analysis [28]. An alternative option are Gabor filters [10, 45] which have optimal time-frequency resolution.

Mel Frequency Cepstral Coefficients–MFCC: The MFCC are the most commonly used feature set for ASR applications. They were introduced by Davis and Mermelstein [9]. Cepstrum analysis can enable the separation of convolved signals. In the linear source-filter model such convolved signals are the source excitation signal and the impulse response of the vocal tract filter. So, the vocal tract's distortions to the speech signal can be removed.

The wide-spread use of the MFCC is due to the low complexity of the estimation algorithm and their efficiency in ASR tasks. In detail the algorithm consists of the following steps. The magnitude-squared of the Fourier transform is computed and triangular frequency weights are applied. These weights represent the effects of the peripheral auditory frequency resolution. Then, the logarithmic outputs of the filterbank are used for the estimation of the cepstrum of the signal. Finally, the feature vectors are estimated with the discrete cosine transform in order to reduce the dimensionality and decorrelate the vector components. These features are smoothed out by dropping the higher-order cepstra coefficients.

Even though MFCC are the most common features for ASR tasks, they appear to have major disadvantages. At first, as most HMMs use Gaussian distributions with diagonal covariance matrices, they cannot benefit from a cepstral liftering, since any multiplying factor that is applied to the observations does not affect the exponent calculation. Second, MFCC are easily affected by common frequency-localized random perturbations which have hardly any effect on human speech communication. Finally, robust feature extraction should process at least about syllable-length (around 200-500 ms) spans of the speech signal [23], in order to extract reliable information for classification of the phonemes.

Another different filterbank approach was proposed by Hermansky [24] and Boulard [7]. They examined Fletcher's proposal [13] to divide the speech spectrum into a number of frequency subbands and extracted spectral features from each of these. However, the recognition/classification task is done independently in each one of the bands by estimating the conditional probabilities for each band. Then, these

estimates are merged in order to give the final output feature set. The merging is done by a multi-layer perceptron (MLP) trained on the same training data as the HMM-based classifiers. The input feature set is the power spectrum values obtained after the PLP critical band filtering, the compression by a cubic-root function and loudness equalization. These features showed a relative improvement of about 50% (compared to the MFCC) in the presence of frequency selective additive noises which corrupted only some of the frequency bands. On the other hand, they were ineffective for noises that corrupted the whole speech spectrum.

Subband Spectral Centroids: These features have been introduced by Paliwal et al, [15]. They can be considered as histograms of the spectrum energies distributed among nonlinearly-placed bins. They show properties similar to those of the distributions of the formant frequencies and they appear to be quite robust to noise. They can be used as a supplementary feature set to cepstral features. Note that the conventional feature sets like the MFCC utilize only amplitude information from the speech power spectrum while the proposed features utilize frequency information, too. It should be stated, though, that these features failed to show significant improvement at the recognition rates when compared to the MFCC.

2.3 Modulations

2.3.1 Short-term Modulations

The linear model of speech makes the assumption that resonances (and the center frequencies of the formants) remain constant for relatively short amounts of time. A nonlinear model proposes that these resonances are not constant but can fluctuate around their center frequency and can be modeled as a sum of AM-FM signals [37]. Short-term modulation features attempt to quantify these fluctuations and capture the temporal and dynamic nature of the speech resonances. The improvement of the recognition rates supports the accuracy of such a nonlinear model [10]. These features are used to enhance the classic cepstrum-based features as an augmented feature set for ASR applications and they show robustness in noisy speech signals due to the use of the filterbank and the Energy Separation Algorithm (ESA).

Alternatively, other algorithms have been proposed to obtain the instantaneous amplitude and frequency signals from the bandpassed speech. Such approaches use Kalman filtering [41] and the Hilbert transform [44]. The first approach has high computational complexity and cannot be used for real-time applications. The latter one has poor temporal resolution and rapid changes are smoothed out.

Finally, some experimentation has been done with merging the source-filter model with the nonlinear model of the resonances for the estimation of MFCC-like features. More specifically, the square amplitude of the bandpassed speech signals is replaced by the nonlinear Teager energy [29] in the standard estimation algorithm of the MFCC feature set. The correct phoneme recognition rates have shown marginal differences for clean speech signals when compared to the MFCC corresponding rates. This is due to the post-processing smoothing effect of the MFCC (i.e. cutting off the higher-order cepstra coefficients). However, the Teager energy-based MFCC features seem to be smoother and more robust for noisy signals and yield improved results.

Frequency Modulation Features: A mel-spaced Gabor filterbank of 6-filters is used in order to bandpass the speech signals. These signals are demodulated using the ESA and the instantaneous frequency and amplitude signals are yielded. Then, the 1^{st} and 2^{nd} moments of the instantaneous frequencys are estimated. The *Frequency Modulation Percentages* (FMP) are the ratio of the second over the first moment of these signals [10]. These spectral moments have been tested as input feature sets for various ASR tasks yielding improved results. For the TIMIT phoneme recognition task the correct phoneme accuracy rates are 40% for the FMPs compared to 55% for the MFCC using though half the vector length of the MFCC. In the AURORA-3 database word recognition task, the relative improvement is 16% compared with the auditory features [8] and 50% when compared with the MFCC [10].

Amplitude Modulation Features: Other nonlinear feature sets have been proposed taking under

consideration the amplitude modulated (AM) part of the nonlinear speech model. The algorithm described above (the filterbank and the demodulation algorithm) has been used to estimate the instantaneous amplitude signals (absolute envelopes) of the bandpassed speech signals. These envelopes are modulated by lowpass signals containing the linguistic information, [10]. The proposed feature set parametrizes the modulating signals and yields their statistics (their 1^{st} and 2^{nd} spectral moments). This feature set shows a statistically important improvement compared to the baseline accuracies of the MFCC. Recent experiments on these features indicate that they are noise invariant, mainly due to their lowpass nature. Namely the instantaneous amplitudes are lowpass signals and, concequently, more robust in noise. Their estimates appear to be very smooth, in terms of spikes and discontinuities, even at low SNR. It has been shown that they contain significant amount of information concerning both the speaker and the linguistic content of the speech signals [46].

The instantaneous amplitude signals, and their corresponding modulating signals, have a very slow temporal evolution. This property is exploited from another viewpoint by research in long-term modulations i.e. the *Modulation Spectrogram*. The short and long-term modulations are two different concepts of the speech production mechanism. The short-term modulations are studied in time-windows up to 10-30ms in order to capture the micro-details (very rapid changes) of the speech signals. On the contrary, long-term modulations examine the temporal evolution of the speech energy and the corresponding time-windows are in the range of 200-500ms.

2.3.2 Long-term Modulations

Early experiments [12, 11, 26] on the perceptual ability of the human auditory system have shown that slow temporal modulations differ as far as their relative importance on different frequencies is concerned. In detail, speech intelligibility is not affected by low-pass filtering below 16 Hz, or high-pass filtering above 4 Hz. Furthermore, intelligibility in noise depends on the integrity of the modulation spectrum in the range between 2 and 8 Hz, on the global shape of the spectral envelope and not so much on its fine details [31]. Finally, the duration of the dominant component (around 4 Hz) is related to the average duration of syllables.

Typical short-time feature extraction methods (filterbank energies, LPC, Cepstrum, MFCC, PLP) form a representation of the spectral envelope of the signal framewise. This has the drawback of being sensitive to background noise. For example at particular frequency components, part of the signal that lies 100ms outside a certain phonetic labeled segment may still carry information relevant to the classification of the given phoneme [55].

The relative importance of the different frequencies of the modulation spectrum is supported in terms of recognition experiments by the different contributions of spectrum components. When the lowest frequency band is removed (cutoff frequency of 1Hz) the accuracy is increased to 93.6% compared to 86% for the unfiltered modulation spectrum. The relative contributions of the various bands of the modulation spectrum, as far as different features are concerned (MFCC, PLP), do not show major differences. The subband that has the maximum contribution is in the range of 2-4Hz of the spectrum. For filterbank related features [31] the more important contribution is in the subband of 4-8Hz and gets more affected by convolutional noise than MFCC and PLP.

The *Dynamic Cepstral Coefficients* method [14] attepmts to incorporate long-term temporal information. These coefficients are computed by first- and second-order orthogonal polynomial expansions of feature time trajectories, referred to as “delta and acceleration coefficients”. They have become a standard method followed by every ASR system and are robust to slowly varying convolution distortions. Alternatively, in the method of *Cepstral Mean Normalization* the long-term average is substracted from the logarithmic speech spectrum and convolutive noise is suppressed.

An alternative to the DC component removal (i.e. Cepstral Mean Normalization), is to use a high-pass filter. In Relative Spectral Processing (RASTA) [21, 22] the modulation frequency components that do not belong to the range from 1 to 12 Hz are filtered out. Thus, this method suppresses the slowly

varying convolutive distortions and attenuates the spectral components that vary more rapidly than the typical rate of change of speech.

Relative Spectra Processing–RASTA: RASTA processing has fundamental relations to both the temporal properties of hearing and the equalization of speech [21]. It achieves a broader, than the delta features, pass-band by adding a spectral pole, and allows the preservation of the linguistic content. RASTA band-pass filtering is applied either on the logarithmic spectrum or on a nonlinearly compressed spectrum and consists of filters with a sharp spectral zero at the zero modulation frequency. The moving average (MA) part of the RASTA filters is derived from the delta features. The spectral pole of the autoregressive (AR) part is obtained through experimentation and determines the high-pass cut-off frequency. The RASTA algorithm consists of the following steps. At first, the critical-band power spectrum is computed. Then, the spectral amplitude is transformed through a compressing static nonlinearity, and the time trajectories of each transformed spectral component are filtered. The filtered speech representation is transformed through an expanding static nonlinearity and is multiplied by the equal loudness curve raised to the power 0.33 in order to simulate the power law of hearing. Finally, an all-pole model of the resulting spectrum is computed.

Several variations have been proposed using a different nonlinear spectral domain than the logarithmic one like the J-RASTA and the Lin-Log RASTA algorithms. The use of such variations improves the correct recognition rates because they simulate more realistically the physiological hearing processes.

Many experiments have been performed in order to examine and compare the RASTA features to other analysis schemes such as the PLP and the PLP+Cepstral Mean Removal. Both logarithmic RASTA and cepstral mean removal improve the recognition rates for convolutional noise. However, PLP, logarithmic RASTA and cepstral mean removal all degrade severely in additive noise. Lin-Log RASTA with a linear mapping shows good robustness over both convolutional and additive noise. While cepstral mean subtraction performed better for purely convolutional noise, it was not as effective as the Lin-Log RASTA approach when additive noise was present.

Temporal Patterns–TRAP: This method was introduced by Hermansky et al. [23]. Conventional features in ASR describe the short-term speech properties. On the other hand, the TRAP features describe likelihoods of sub-word classes at a given time instant, derived from temporal trajectories of band-limited spectral densities in the vicinity of the given time instant.

Coding of linguistic information in a single short-term spectral frame of speech appears to be very complex. A single frame of such a short-term spectrum does not contain all the necessary information for the decoding scheme as the neighboring speech sounds influence the short-term spectrum of the current one. The mechanical inertia of human speech production organs results in spreading the linguistic information in time. At any given time at least 3-5 phonemes interact. This introduces high within-phoneme variability of the spectral envelope. ASR systems attempt to classify phonemes from individual slices of the short-term spectrum and need to deal with this within-class variability, even though experiments show that human listeners are not affected by such phenomena.

Such ASR systems expect feature vectors of uncorrelated and normally distributed features every 10 ms. So, a process is needed that is capable of examining long spans of speech within various frequency bands and deliver every 10 ms uncorrelated and normally distributed features. The proposed algorithm TRAP-TANDEM is such a module. The tandem submodule is an hierarchical tree-based structure that splits speech into different sound classes e.g. voiced, unvoiced, silence, etc.

This processing scheme is capable of examining relatively long spans of the speech signal within various frequency bands. It uses MLP (Multi-Layer Perceptron) to provide nonlinear mapping from temporal trajectories to phoneme likelihoods. The *TRAP* processing uses relatively long time windows (500-1000 ms) and frequency localized (1-3 Barks) overlapping time-frequency regions of the signal. The *TANDEM* algorithm refers to a way of converting the frequency-localized evidence to features for the HMM-based ASR systems.

The time-frequency spectral density plane is estimated using the front-end taken from the PLP analysis. It employs the short-time spectral analysis of the speech signal using a Bark-spaced filterbank. The

input to the TRAP estimator consists of 1-3 time trajectories of critical-band energies. The individual trajectories are concatenated to form a longer input vector, and finally, PCA is introduced in order to reduce the vector's dimensionality.

TRAP estimator delivers vectors of posterior probabilities sub-word acoustic events, each estimated at an individual frequency band. The events targeted are the phonemes clustered into 6 broad phonetic classes, and separate estimators are trained for each frequency region of interest.

The TANDEM part derives a vector of posterior probabilities of sub-word speech events for every speech frame from the evidence presented to its input. An MLP is used in order to optimally cluster the input vectors. and estimate the posterior probabilities of the individual classes. These probabilities are post-processed by a static nonlinearity in order to match the gaussian probability distributions, and whitened by the KL transform derived by the training data.

The events targeted by the TRAP estimators do not need to be the same as those targeted by the TANDEM estimator. Also, TRAP estimators can be trained on different databases than the databases used in training the TANDEM estimator. Note that both the TRAP and TANDEM estimators are nonlinear feed-forward MLP discriminative classifiers.

So far, the TRAP-TANDEM features have been found more useful in combination with the conventional spectrum-based features like PLP and MFCC, where they brought more than 10% relative improvement (for the DARPA EARS program). Nowdays, the performance of the TRAP-TANDEM stand-alone features is becoming comparable with the traditional approaches. For example, for the OGI Numbers task they yield the same (5%) word error rate as the best system using the PLP+Delta+DDelta features. Finally, for the TIMIT database task the TRAP-based features gave 10% relative improvement in the phoneme error rates compared to the MFCC.

2.4 Auditory-based Features

The human auditory system is a biological apparatus with ideal performance, especially in noisy environments. Various ASR approaches incorporate characteristics of this system. The adaption of physiologically based methods for spectral analysis [16] is such an approach. The physiological model of the auditory system can be categorized into the areas of outer, middle and inner ear [48]. The cochlea and the basilar membrane, both located in the inner ear, can be modeled as a mechanical realization of a bank of filters. Along the basilar membrane are distributed the Inner Hair Cells (IHC), which sense mechanical vibrations and convert them into firing of the connected nerve fibers which in turn emit neural impulses to the auditory nerve.

Inspired by the above ideas, the **Ensemble Interval Histogram** (EIH) model is constructed by a bank of 'cochlear' filters followed by an array of level crossing detectors that model the motion to neural conversion. The probability distributions of the level crossings are summed up for each cochlear filter resulting on the ensemble interval histogram. Front-ends that use ideas of this approach have shown comparable recognition rates to common spectrum-based features. Moreover, they are characterized by increased noise resistance for lower SNR's [17].

Lateral inhibition is another characteristic that has been introduced into periphery models. This is defined as the suppression of the activity of nerve fibers on the basilar membrane caused by the activity of adjacent fibres. It accounts for the phenomenon caused when two tones of different amplitude are similar in frequency, leading to an inhibition in the perception of the weaker one. The aforementioned phenomenon has been used to improve noise robustness by convolving a frequency dependent lateral inhibition function with noisy speech [56]. Since narrowband SNR is higher on spectral peaks, by emphasising these areas and attenuating spectral valleys, the signal's SNR increases.

The **Joint Synchrony/Mean-Rate** model [50, 51] captures the essential features extracted by the cochlea in response to sound pressure waves. It includes parts that deal with peripheral transformations occurring in the early stages of the hearing process. These parts attempt to extract information relevant to perception, such as formants, and enhance sharpness of onset and offset of different speech segments. In

detail, the speech signal is first pre-filtered through a set of four complex zero pairs to eliminate the very high and very low frequency components. Then it passes through a 40-channel critical-band linear filter bank whose single channels were designed in order to fit physiological data. Next, the hair cell synapse model is intended to capture prominent features of the transformation from basilar membrane vibration, represented by the outputs of the filter bank, to probabilistic response properties of auditory nerve fibers. The outputs of this stage represent the probability of firing as a function of time for a set of similar fibers acting as a group. The two output models that follow are the Generalized Synchrony Detector (GSD) and the Envelope Detector (ED). GSD which implements the known "phase-locking" property of nerve fibers, is designed with the aim of enhancing spectral peaks due to vocal tract resonances. ED computes the envelope of the signals at the output of the previous stage of the model and is important for capturing the very rapidly changing dynamic nature of speech.

An important type of filter that has been proposed for auditory processing is the **gammatone** function [28]. This has been shown to describe impulse-response data gathered physiologically from primary auditory filters in the cat. The gammachirp is constructed by adding a frequency modulation term to the gammatone function. This function has minimal uncertainty in joint time/scale representation. The gammachirp auditory filter is the real part of the analytic gammachirp function, has an asymmetric amplitude characteristic and provides an excellent fit to human masking data.

Auditory peripheral modeling is another area that incorporates auditory characteristics. These include critical band filtering, loudness curve properties, nonlinear energy compression, haircell modeling and short-time adaptation. By the use of such models there are improvements in the temporal localization and speech detectability in degraded environments, resulting in increased system robustness to noise.

Perceptual linear prediction (PLP) is a variant of Linear Prediction Coding (LPC) which incorporates auditory peripheral knowledge [19, 20]. The main characteristics for estimating the audible spectrum are realized by adding critical band integration, equal-loudness pre-emphasis and intensity to loudness compression. More specifically, the method considers the short-term power spectrum and convolves it with a critical-band masking pattern. Then, the critical band is resampled at about one Bark scale intervals. A pre-emphasis operation is performed with a fixed equal loudness curve and finally the resulting spectrum is compressed with a cubic root nonlinearity. The output low-order all-pole model is consistent with phenomena observed in human speech perception. It simulates the properties of the auditory system resulting in parameters compatible with LPC. The main advantage of PLP is the reduction of the order of the model (e.g. 5 coefficients vs. 15 for LPC).

2.5 Fractal-based Features

One of the latest approaches in speech analysis are the nonlinear/fractal methods. These diverge from the standard linear source-filter approach in order to explore nonlinear characteristics of the speech production system. They are based on tools that lie in the areas of fractals and dynamical systems. Their motivation stems from the observations of aerodynamic phenomena in speech production [52, 30]. Specifically, airflow separation, unstable air jet, oscillations between the walls and vortices are phenomena encountered in many speech sounds and lead to turbulent flow. Especially fricatives, plosives and vowels uttered with some speaker-dependent aspiration contain various amounts of turbulence. Moreover, the presence of vortices could result in additional acoustic sources. The initial significant contributions [52, 30] are further supported by acoustic and aerodynamic analysis of mechanical models [5, 25]. On the other hand it has been conjectured that geometrical structures in turbulence can be modeled using fractals [35]. Difference equation, oscillator and prediction nonlinear models were among the early works in the area [47, 33, 54]. Speech processing techniques that have been inspired by fractals have been introduced in [36, 38]. These measure the roughness of the signal in multiple scales as a quantification of the geometrical complexity of the underlying signal. Their application as short-time features in ASR experiments has shown significant improvement of 12%-18% error reduction in the tough recognition task over the E-set ISOLET database [38].

Recently various approaches [18, 34, 40, 53, 6, 49] apply such fractal-based measurements on reconstructed multidimensional phase spaces instead of the one-dimensional signal space. They argue that the multidimensional reconstructed space is closer to the true speech production dynamics compared to the one dimensional speech signal. The latter can be seen as a collapsed projection from a higher dimensional space. The analysis is carried out by computing invariants of the multidimensional signals, which are the fractal dimensions and Lyapunov exponents [32], [4], [43]. Generalized dimensions [42],[3] and multifractal spectrum [1] are alternative representations of the underlying geometrical complexity. Special cases include the standard fractal dimension (box-counting, Minkowski-Boulingand dimension), correlation dimension and information dimension. It should be noted that this field is not fully developped yet because the observed phenomena are neither completelly understood nor directly related to the various approaches and models reported. Moreover, a suitable way to integrate such analysis in ASR systems is not a simple task and only preliminary results have been reported [38].

2.6 Discussion

In this report we have presented briefly the main trends for robust feature extraction techniques in ASR systems. Feature extraction methods can be categorized into overlapping classes that share a number of common ideas. The most common ideas are related to filterbank processing, features inspired by the physiology of the auditory system, features utilizing perceptual knowledge, or inspired by phenomena that occur during speech production (e.g. modulations). A review of the proposed features for ASR systems indicates that cepstral analysis and the *MFCC* [9] features have become one of the most common approaches. A popular alternative are the *PLP* [20] or related features that are based on knowledge of the human auditory peripheral system. Finally, nonlinear speech processing techniques (e.g. modulations, fractals) have started to gain momentum. Many techniques share the concept of short-time processing. However, recently there have been introduced alternative methods e.g. *RASTA* [21] ,*TRAP* [23] that filter out parts of the modulation spectrum or process frames that span longer time intervals. There are not direct comparisons for every proposed feature set, but implicit conclusions may be assumed by considering their absolute recognition results.

Concluding, although research in this area has been active for many decades, robustness is still a key issue that should be considered. Thus more effort should be placed in order to accomplish satisfactory performance in adverse acoustic environments.

2.7 Useful URLs

Speech at Carnegie Mellon University:
`http://www.speech.cs.cmu.edu`

IDIAP Speech Processing Group:
`http://old-www.idiap.ch/speech/speechNF.html`

Center for spoken language understanding at Oregon Health and Science Univercity:
`http://cslu.cse.ogi.edu`

Center for Spoken language research at Univercity of Colorado:
`http://cslr.colorado.edu`

Spoken Language Systems at MIT Laboratory for Computer Science:
`http://www.sls.csail.mit.edu/sls/sls-blue-nospec.html`

The International Computer Science Institute Speech Group at Berkeley:
`http://www.icsi.berkeley.edu/Speech/`

Speech processing and Auditory perception laboratory at UCLA:

http://www.icsl.ucla.edu/~spapl

Various links of research groups:
http://mambo.ucsc.edu/psl/speech.html

Bibliography

[1] O. Adeyemi and F. G. Boudreaux-Bartels. Improved accuracy in the singularity spectrum of multifractal chaotic time series. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-97,* Munich, Germany, 1997.

[2] J. B. Allen. How do humans process and recognize speech? *IEEE Trans. Speech Audio Processing*, 4(2):567–577, 1994.

[3] Y. Ashkenazy. The use of generalized information dimension in measuring fractal dimension of time series. *Physica A*, 271(3-4):427–447, 1999.

[4] M. Banbrook, S. McLaughlin, and I. Mann. Speech characterization and synthesis by nonlinear methods. *IEEE Trans. Speech Audio Processing*, 7:1–17, 1999.

[5] A. Barney, C. H. Shadle, and P. O. A. L. Davies. Fluid flow in a dynamic mechanical model of the vocal folds and tract. i. measurements and theory. *J. Acoust. Soc. Am.*, 105(1):444–455, 1999.

[6] H.-P. Bernhard and G. Kubin. Speech production and chaos. In *XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, France, Aug 19-24, 1991.

[7] H. Bourlard and S. Dupont. A new asr approach based on independent processing and recombination of partial frequency bands. In *Proc. International Conference on Speech and Language Processing, ICSLP-96*, pages 426–429, Philadelphia, USA, 1996.

[8] J. Chen, D. Dimitriadis, H. Jiang, Q. Li, T. A. Myrvoll, O. Siohan, and F. K. Soong. Bell labs approach to aurora evaluation on connected digit recognition. In *Proc. International Conference on Speech and Language Processing, ICSLP-02*, pages 462–465, Denver, CO, USA, September 2002.

[9] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, 28(4):357–366, 1980.

[10] D. Dimitriadis and P. Maragos. Robust energy demodulation based on continuous models with application to speech recognition. In *Proc. European Conference on Speech Communication and Technology, Eurospeech-03*, Geneva, Switzerland, September 2003.

[11] R. Drullman, J. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.*, 95(5):2670–2680, 1994.

[12] R. Drullman, J. Festen, and R. Plomp. Effect of temporal smearing on speech reception. *J. Acoust. Soc. Am.*, 95(2), 1994.

[13] H. Fletcher. *Speech and Hearing in Communication*. Krieger, New York, 1953.

[14] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust., Speech, Signal Processing*, 29(2):254–272, 1981.

[15] B. Gajic and K. K. Paliwal. Robust feature extraction using subband spectral centroid histograms. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-01*, volume 1, pages 85–88, 2001.

[16] O. Ghitza. Auditory nerve representation as a front-end in a noisy environment. *Computer Speech and Language*, 2(1):109–130, 1987.

[17] O. Ghitza. Auditory nerve representation as a basis for speech processing. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 453–486. Marcel Dekker, New York, 1992.

[18] S. Haykin and J. Principe. Making sense of a complex world. *IEEE Signal Processing Magazine*, page 6681, May 1998.

[19] H. Hermansky. An efficient speaker independent automatic speech recognition by simulation of some properties of human auditory perception. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-87*, pages 1156–1162, 1987.

[20] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752, 1990.

[21] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Trans. Speech Audio Processing*, 2(4):578–589, 1994.

[22] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In *Proc. European Conference on Speech Communication and Technology, Eurospeech-91*, pages 578–589, 1991.

[23] H. Hermansky and S. Sharma. Traps - classifiers of temporal patterns. In *Proc. International Conference on Speech and Language Processing, ICSLP-98*, 1998.

[24] H. Hermansky, S. Tibrewala, and M. Pavel. Towards asr on partially corrupted speech. In *Proc. International Conference on Speech and Language Processing, ICSLP-96*, pages 462–465, Philadelphia, USA, 1996.

[25] H. Herzel, D. Berry, I. Titze, and I. Steinecke. Nonlinear dynamics of the voice: Signal analysis and biomechanical modeling. *CHAOS*, 5(1):30–34, 1995.

[26] T. Houtgast and H. J. M. Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.*, 77(3):1069–1077, 1985.

[27] M.J. Hunt and C. Lefebvre. Speech recognition using a cochlear model. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-86*, pages 1979–1982, 1986.

[28] T. Irino and R. D. Patterson. A time-domain, level-dependent auditory filter: The gammachirp. *J. Acoust. Soc. Am.*, 101:412–419, 1997.

[29] F. Jabloun, A. E. Cetin, and E. Erzin. Teager energy based feature parameters for speech recognition in car noise. *IEEE Signal Processing Lett.*, 6(10), 1999.

[30] J. F. Kaiser. Some observations on vocal tract operation from a fluid flow point of view. In I. R. Titze and R. C. Scherer, editors, *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, pages 358–386. Denver Center for Performing Arts, Denver, CO, 1983.

[31] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel. On the importance of various modulation frequencies for speech recognition. In *Proc. European Conference on Speech Communication and Technology, Eurospeech-97*, pages 1079–1082, Rhodes, Greece, 1997.

[32] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. *IEEE Trans. Acoust., Speech, Signal Processing*, to be published.

[33] G. Kubin and W. B. Kleijn. Time-scale modification of speech based on a nonlinear oscillator model. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-94*, 1994.

[34] A. Kumar and S. K. Mullick. Nonlinear dynamical analysis of speech. *J. Acoust. Soc. Am.*, 100(1):615–629, 1996.

[35] B. Mandelbrot. *The Fractal Geometry of Nature*. Freeman, NY, 1982.

[36] P. Maragos. Fractal aspects of speech signals: Dimension and interpolation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-91*, 1991.

[37] P. Maragos, J. F. Kaiser, and T. F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Trans. on Signal Processing*, 41(10):3024–3051, October 1993.

[38] P. Maragos and A. Potamianos. Fractal dimensios of speech sounds: Computation and application to automatic speech recognition. *J. Acoust. Soc. Am.*, 105(3):1925–1932, 1999.

[39] C. Nadeu, D. Macho, and J. Hernardo. Time and frequency filtering of filterbank energies for robust hmm speech recognition. *Speech Communication*, 34:93–114, 2001.

[40] S. Narayanan and A. Alwan. A nonlinear dynamical systems analysis of fricative consonants. *J. Acoust. Soc. Am.*, 97(4):2511–2524, 1995.

[41] W.-C. Pai and P. C. Doerschuk. Statistical am-fm models, extended kalman filter demodulation, cramer-rao bounds, and speech analysis. *IEEE Trans. on Signal Processing*, 48(8):2300–2313, August 2000.

[42] V. Pitsikalis, I. Kokkinos, and P. Maragos. Nonlinear analysis of speech signals: Generalized dimensions and lyapunov exponents. In *Proc. European Conference on Speech Communication and Technology, Eurospeech-03*, Geneva, Switzerland, September 2003.

[43] V. Pitsikalis and P. Maragos. Speech analysis and feature extraction using chaotic models. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-02*, Orlando, USA, May 2002.

[44] A. Potamianos and P. Maragos. A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation. *Signal Processing*, 37:95–120, May 1994.

[45] A. Potamianos and P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *J. Acoust. Soc. Am.*, 99(6):3795–3806, June 1996.

[46] A. Potamianos and P. Maragos. Speech analysis and synthesis using an am-fm modulation model. *Speech Communication*, 28:195–209, 1999.

[47] T. F. Quatieri and E. M. Hofstetter. Short-time signal representation by nonlinear difference equations. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'90, Albuquerque*, April 1990.

[48] L. R. Rabiner and B.H.Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[49] S. Sabanal and M. Nakagawa. The fractal properties of vocal sounds and their application in the speech recognition model. *Chaos, Solitons and Fractals*, 7 No11:1825–1843, 1996.

[50] S. Seneff. Pitch and spectral estimation of speech based on an auditory synchrony model. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-84*, pages 3621–3624, 1984.

[51] S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16(1):57–76, 1988.

[52] H. M. Teager and S. M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech Production and Speech Modelling W.J. Hardcastle & Marchal, Eds., NATO ASI Series D*, volume 55, 1989.

[53] N. Tishby. A dynamical systems approach to speech processing. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-90*, 1990.

[54] B. Townshend. Nonlinear prediction of speech signals. *IEEE Trans. Acoust., Speech, Signal Processing*, 1990.

[55] H. H. Yang, S. J. Van Vuuren, and H. Hermansky. Relevancy of time-frequency features for phonetic classification measured by mutual information. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-99*, 1999.

[56] Y.M.Cheng and D.O'Shaughnessy. Speech enhancement based conceptually on auditory evidence. *IEEE Trans. Acoust., Speech, Signal Processing*, 39(9):1943–1954, 1991.

Chapter 3

Speech analysis

Y. Laprie
INRIA-Parole, France

3.1 Description of the problem

Among the variety of acoustic signals the ear is exposed to, speech is one of the few for which an approximated production model is available. All the speech signals thus share common characteristics that are of interest for various areas in automatic speech processing: speech synthesis, automatic speech recognition, speech synthesis, hearing aids, telecommunications...

The production of a speech signal $s(n)$ can be approximated as the convolution of an excitation signal $e(n)$ by a filter corresponding to the vocal tract $h(n)$: $s(n) = e(n) * h(n)$. In the spectral domain this equation becomes $S(\omega) = E(\omega) \times H(\omega)$.

Speech analysis aims at finding contributions of the excitation signal (noise in the case of unvoiced sounds or periodic signal for voiced sounds) and of the vocal tract filter, and separating these two contributions.

The excitation signal is produced by either the vibration of vocal folds (periodic excitation) or the turbulent flow of air somewhere in the vocal tract (noise excitation).

The filter corresponding to the vocal tract depends its geometrical shape and thus depends on the position of speech articulators, i.e. the lower jaw, position and shape of the tongue body, position of the tongue apex, aperture and rounding of lips, larynx and velum position.

The three main areas of research in speech analysis are **(i)** spectral analysis, **(ii)** the determination of the fundamental frequency and **(iii)** automatic formant tracking.

3.2 State of the art

3.2.1 Spectral analysis

The main objective of spectral analysis is to get a relevant spectrum of the vocal tract filter. The main challenge is to get an information as independent as possible of the excitation. The spectrum of a voiced excitation is a spectrum of regularly spaced (the fundamental frequency F0) lines. The vocal tract spectrum is thus "sampled" by F0. The higher F0, the less the precision of the vocal tract spectrum. This means that the vocal tract spectrum is not well approximated for high F0 voices, i.e. for female speakers and children.

The first tool for spectral estimation is the well know Fourier transform. The size of the analysing window influences the frequency smoothing. When the time window is small (approximately 4 ms) compared to the F0 period the smoothing is strong and resonance frequencies of the vocal tract can be seen. However, the spectrum depends on the location of the window with respect to F0 periods and this solution is only used to display speech spectrograms used by phoneticians. When the time window

is long (approximately 32 ms) compared to the F0 periods the smoothing is weak and thus, harmonics (multiples of F0) are visible. Fourier analysis is the basic tool for spectral analysis of speech. One of the difficulties is the choice of the size and position or the analysing window with respect to the periods of the fundamental frequency. There exists some reassignment methods that reduce the effect of the window location by moving the spectral energy where it should appear [11].

Beside Fourier transform there are two main families of spectral analyses in speech processing. The first is that of linear prediction methods that correspond to the assumption of an all pole model. The idea is to approximate the speech sample $s(n)$ by $\sum_{k=1}^{k=p} a_k s(n-k)$. Parameters a_k are obtained by minimising the squared error over an analysing window. The spectrum can be easily calculated from these coefficients. The advantage of this method is its low computation cost. Its main disadvantage is that the underlying all pole hypothesis is only valid for oral vowels and not for nasal vowels and consonants. There are a number of derived methods, the selective linear prediction for instance, that enable the analysis to be applied over a limited spectral region [10].

The second family is that of cepstral smoothing. The principle is to eliminate the contribution of the excitation in a Fourier spectrum calculated over a rather long window (approximately 32 ms). The underlying idea is to compute the inverse Fourier transform of the spectrum, called cepstrum, to isolate the contribution of harmonics. This contribution appears as an isolated peak that can be easily filtered. An extra Fourier transform gives the smoothed spectrum (see [12] for a more detailed presentation). Derived from this idea Davis and Mermelstein [2] proposed Mel cepstra that are calculated from the energy vector computed over a Mel[1] filter bank after the Fourier transform. This enables a more concise representation of speech spectra and removes speaker variability to some extent. These Mel cepstral coefficients are widely used as input spectral vectors in automatic speech recognition. There are interesting methods derived from cepstral analysis. The spectral envelope method [7] is an iterative version of the standard cepstral analysis that enables a good energy approximation in the vicinity of harmonics. This gives a better approximation and a more steady estimation of spectral peaks. The discrete cepstral analysis [3] approximates a number of spectral points by a sum of cosinusoids. The spectral points have to be chosen carefully to represent relevant spectral information. One therefore chooses harmonics or other spectral peaks.

Despite their theoretical interest wavelets are not often used.

3.3 Determination of the fundamental frequency

The fundamental frequency is the frequency of vocal fold vibration. When vocal folds vibrate, the vocal tract is excited by a periodic signal which gives rise to voiced sounds. The fundamental frequency, called F0, often improperly called pitch which is related to perception, plays a central role in speech analysis. Indeed, the fundamental frequency the is the prosody parameter that gives intonation and, as explained above, has a major impact of the shape of the speech spectrum. Its determination thus has received considerable attention. Furthermore, the fundamental frequency is very important within the framework of speech coding and synthesis. There are basically two kinds of determination method: (i) methods that operate in the time domain as the famous autocorrelation method [5, 12], and (ii) methods that operate in the frequency domain as the cepstrum or spectral comb methods. The difficulties lie in the false determination of double F0 or half F0 values, and in the voicing decision, i.e. how to decide whether a speech window corresponds to voiced or unvoiced speech. These problems and the processing of noisy signal are the current challenges.

[1] The Mel frequency is a non linear frequency scale that approximates the frequency resolution of the ear.

3.4 Automatic formant tracking

As explained in the introduction one of the objectives in speech analysis is to find spectral information related to the vocal tract filter. Formants are spectral peaks that correspond to the resonance frequencies of the vocal tract. As formants directly derive from the geometrical shape of the vocal tract, they may be exploited to recover the place of articulation and thus identify sounds pronounced, especially vowels and other vocalic sounds. Formant tracks are utilised to pilot formant synthesisers [6], to study coarticulation effects, vowel perception, articulatory phenomena and in some rare cases to provide a speech recognition with additional data [4].

Given the potential interest of formant data numerous works have been dedicated to the design of automatic formant tracking algorithms. Nature and complexity of the problem explain the success of dynamic programming algorithms [13, 14]. The first steps of these algorithms is the extraction of formant candidates at each frame of the speech signal. The second stage is dynamic programming that utilises the evaluation of transition costs between two frames. Other algorithms aim at explaining the acoustic signal [1] or the spectrogram energy. In [9, 8] we showed how active curves could be used to track formants. The underlying idea is to deform initial rough estimates of formants under the influence of the spectrogram to get regular tracks close to lines of spectral maxima which are potential formants.

3.5 Perspectives

Despite of constant efforts, automatic formant tracking remains an open challenge. This challenge is all the more important since good formant estimates could be exploited in various areas of automatic speech processing: speech recognition, synthesis, speaker identification... As formants are closely related to speech production, analysis by synthesis methods are the most promising approach. Moreover, progress in speech production in terms of talking heads and speaker adaptation could provide additional constraints to improve results.

In the domain of F0 determination the robustness is still an open challenge, especially when performances are compared against those of human listeners who are able to detect speech even in a strong ambiant noise. Improvement of F0 determination techniques is probably closely linked to the development of new spectral analyses that achieve a better precision in the localization of energy. Another potential source of improvement is a better cooperation with psychoacoustics that focuses on the human perception of acoustic signals and investigates the reasons why the ear is far better than early processing of speech.

Bibliography

[1] I. Bazzi, A. Acero, and L. Deng. An expectation maximization approach for formant tracking using a parameter-free non-linear predictor. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Hong-Kong, May 2003.

[2] S. B. Davis and P. Mermelstein. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-28(4):357–366, August 1980.

[3] T. Gallas and X. Rodet. Generalized functionnal approximation for source-filter system modelling. In *Proceedings of European Conference on Speech Technology*, Genova, Italy, September, 1991.

[4] Philip N. Garner and Wendy J. Holmes. On the robust incorporation of formant features into hidden makov mdels for automatic speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1–4, Seattle, USA, May 1998.

[5] W. J. Hess. *Pitch Determination of Speech Signals - Algorithms and Devices*. Springer Berlin, 1983.

[6] J. N. Holmes. Formant synthetisers: cascade or parallel ? *Speech Communication*, 2:251–273, 1983.

[7] S. Imai and Y. Abe. Spectral envelope extraction by improved cepstral method. *Trans. IECE*, J62-A(4):217–223, 1979 (en japonais).

[8] Y. Laprie. A concurrent curve strategy for formant tracking. In *International Conf. on Spoken Language Processing - ICSLP2004, Jegu, Korea*, October 2004.

[9] Y. Laprie and M.-O. Berger. Cooperation of regularization and speech heuristics to control automatic formant tracking. *Speech Communication*, 19(4):255–270, October 1996.

[10] J.D. Markel and A.H. Gray. Automatic formant trajectory estimation. In *Linear Prediction of Speech*, chapter 7. Springer-Verlag, Berlin Heidelberg New York, 1976.

[11] F. Plante, G. Meyer, and W.A. Ainsworth. *IEEE Transactions on Speech and Audio Processing*, 6(3):282–287, 1998.

[12] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, N.J., 1978.

[13] D. Talkin. Speech formant trajectory estimation using dynamic programming with modulated transition costs. *Journal of the Acoustical Society of America*, S1:S55, March 1987.

[14] K. Xia and C. Epsy-Wilson. A New Strategy of Formant Tracking based on Dynamic Programming. In *International Conf. on Spoken Language Processing - ICSLP2000, Beijing, China*, October 2000.

Chapter 4

Speech analysis with emphasis to speech emotion recognition

C. Kotropoulos
AUTH, Greece

4.1 Introduction

Speech analysis has a variety of sectors, including speech modelling, speech recognition and synthesis, acoustics-phonetics, voice coding and compression, speech enhancement, speaker identification/ classification, emotion recognition, speech processing for the hearing impaired, linguistics, and psychoacoustics. The basic assumption in speech analysis is that speech is short time stationary. From a certain short frame of speech a feature vector must be formulated to be used for future feature processing.

4.2 Speech analysis today

The widely used speech processing technologies are speech compression (compression of the signal for transmission through limited-bandwidth channels), speech synthesis and speech recognition (recognition of voice input by the machine). Speech analysis is involved in many applications for communicating or transmitting information over the wired network, including voice mail, beeper services, ISDN video, FAX, audio and video teleconferencing, etc. Speech compression is considered a saturated field, whereas speech recognition and synthesis are not. Some of the problems in speech recognition are speaker identification and emotion recognition. Emotion recognition classifies speech in categories, which are related to the psychological state of the user. The usual emotion categories are anger, fear, sad, happy, disgust and surprise. The term *"basic emotions"* is widely used without implying that these emotions can be mixed to produce others [2].
Text to speech synthesis is met in many applications which involve human-computer interaction, such as voiced e-mails, sites, and books, that are read to a user. The lack of emotion variability in synthetic speech clearly affects the speech quality and speech rate and makes the resulting narrations to sound monotonous and non-realistic. Several projects have addressed emotional speech synthesis, an example is CHATAKO-AID [4].
Speech emotion recognition would be a great achievement of computer technology. The words "Affective Computing" are often used to describe software and hardware which are adjusted to the human emotional state. A user is not always happy when uses a PC. For example, experiments in MIT labs [15] have revealed that the user acts violently onto the machine when his job is not accomplished. Emotion recognition from speech can improve the quality of service in call center applications. Several project use emotion recognition to accommodate telephone users in call center applications. Automatic dialog

systems with the ability to recognize the emotions of callers and to modify the system response accordingly are designed in [7].
Emotional speech recognition is closely related to automatic speaker verification (ASV). The voice variability related to the speaker emotional state is studied for ASV in EMOVOX [10]. Another project related to ASV is VERIVOX, which explores speech under a variation of conditions including low level psychological stress [11].
The emotion recognition is rather perceptual than subjective. The classification scores of an automatic algorithm must be compared with scores of humans. The human classification-perception rates vary from 55% to 90% according to the number of emotions and the database. The automatic classification scores are also depended to the same factors. Investigations report automatic correct classification scores at the level of 55% for 5 emotions, whereas the random classification achieves 20%.
The techniques for emotion recogition are classified into three categories:

- Classical approaches to classification and discrimination, like Support Vector Machines, Linear Discriminant Analysis, and K- Nearest Neighbors [8].
- Hidden Markov Model (HMM)-based techniques which have been widely used for speech recognition [3].
- Hybrid methods making use of both discriminant classifiers for classification and HMMs for likelihood estimation are reported in subsection.

The classical discrimination and HMM based techniques have almost equal scores. Hybrid methods tend to be better but are not fully investigated yet. More information about emotion recognition and affective computing can be found in the pages of MIT [15] and the Center for Spoken Language Research [12]. A detailed review about emotion recognition can be found in [1].

Bibliography

[1] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, “Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, pp. 33-60, 2003.

[2] P. Eckman, “An argument for basic emotions," *Cognition and Emotion* vol. 6, pp. 169-200, 1992.

[3] J. H. L. Hansen, D. A. Cairns, “ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect enviroments," *Speech Communication*, vol. 16, pp. 391-422, 1995.

[4] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, “A speech synthesis system with emotion for assisting communication," in *Proc. ISCA Workshop on Speech and Emotion 2000 (ITRW)*, pp. 167-172, Belfast, 2000.

[5] S. J. L. Mozziconacci and D. J. Hermes, “Expression of emotion and attitude through temporal speech variations," in *Proc. Sixth Int. Conf. Spoken Language Processing (ICSLP 2000)*, vol. 2, pp. 373-378, Beijing, China, October 2000.

[6] R. Nakatsu, A. Solomides, and N. Tosa, “Emotion recognition and its application to computer agents with spontaneous interactive capabilities," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems (ICMCS ’99)*, vol. 2, pp. 804-808, Florence, Italy, July 1999.

[7] V. A. Petrushin, “Emotion in speech recognition and application to call centers," in *Proc. Artificial Neural Networks In Engineering (ANNIE 99)*, pp. 7-10, St. Louis, USA .

[8] D. Ververidis, C. Kotropoulos, and I. Pitas, “Automatic emotional speech classification," in *Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2004)*, vol. 1, pp. 593-596, Montreal, Canada.

[9] B. D. Womack and J. H. L. Hansen, “N-Channel Hidden Markov Models for combined stressed speech classification and recognition," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 6, pp. 668-677, November 1999.

[10] EMOVOX project, http://www.unige.ch/fapse/emotion/projects/

[11] Verivox project, “Verivox: Voice variability in speaker verification," http://www.speech.kth.se/ iger/verivox.html

[12] The Center for Spoken Language Research (CSLR), CU Kids’ speech corpus, http://cslr.colorado.edu/beginweb/speechcorpora/corpus.html

[13] NATO IST-03 (Formerly RSG. 10) speech under stress web page, http://cslr.colorado.edu/rspl/stress.html

[14] PHYSTA project, http://www.image.ece.ntua.gr/physta/

[15] Affective computing, MIT media laboratory, http://affect.media.mit.edu/

[16] INTERFACE: Multimodal Analysis/Synthesis System for Human Interaction to Virtual and Augmented Environments, http://gps-tsc.upc.es/imatge/_Montse/INTERFACE.html

[17] Dialog-based Human-Technology Interaction by Coordinated Analysis and Generation of Multiple Modalities, http://www.smartkom.org/start_en.html

Chapter 5

State of the art in acoustic-to-articulatory inversion

Y. Laprie
INRIA-Parole, France

5.1 Description of the problem

The acoustic-to-articulatory inversion consists in recovering the vocal tract shape dynamics from the acoustical speech signal, that can possibly be completed by the knowledge of the speaker's face. Estimating the vocal tract shape from the speech signal has received considerable attention because it offers new perspectives for speech processing. Indeed, it would enable knowing how a speech signal has been articulated. This potential knowledge could give rise to a number of breakthroughs in automatic speech processing. For speech coding, this would allow spectral parameters to be replaced by a small number of articulatory parameters[8] that vary slowly with time. In the case of automatic speech recognition the location of critical articulators could be exploited[7] in a view of discarding some acoustical hypotheses. For language acquisition and second language learning this could offer articulatory feedbacks. Lastly, in the domain of phonetics, inversion would enable knowing how sounds were articulated without requiring medical imaging techniques.

Basically, the acoustic-to-articulatory inversion is an acoustical problem and data are therefore formant frequencies, i.e. the resonance frequencies of the vocal tract. However, formants cannot be extracted easily from the speech signal and most of the existing methods thus need to be generalized to accept standard spectral input data (for instance, Mel Frequency Cepstral Coefficients). This represents a first difficulty to solve the inverse problem.

The main difficulty is that acoustic-to-articulatory inversion is an ill-posed problem. There is no one-to-one mapping between acoustic and articulatory domains and there are thus an infinite number of vocal tract shapes that can produce the same formants and thus the same speech signal. Indeed, the problem is under-determined because there are more unknowns than data. Generally the first three formant frequencies are used as data and there are more than six articulatory parameters, for instance seven in the case of the famous Maeda's model [5]. One important issue is thus to add constraints that are both sufficiently restrictive and realistic from a phonetic point of view.

5.2 State of the art

Most of the acoustic-to-articulatory methods rest on an analysis-by-synthesis approach. Indeed, among the variety of acoustical signals the ear is exposed to, speech is one of the few an approximated production model is available for. The synthesis corresponds to the use of an articulatory synthesizer that computes speech spectra or formants from articulatory or geometrical parameters. Adjusting the faithfulness of the

articulatory synthesizer with respect to the static and dynamic characteristics of the human vocal tract allows constraints to be put on the shape of inverse solutions and thus the number of inverse solutions to be reduced. The simplest articulatory models approximate the vocal tract shape geometrically as a set of concatenated uniform tubes (generally between 6 and 8 tubes). Their main weakness is that they are unable to render the vocal tract shape and that the total length of the vocal tract is an extrinsic parameter that has to be calculated independently. More faithful models can be built from medical imaging of the vocal tract. The 2D sagittal articulatory model proposed by Meada [5] was derived from X-ray images and describes the vocal tract through seven deformation modes. More recent models based on MRI images describe the 3D shape of the vocal tract [2] articulatory parameters. The strong noise and lying position imposed by a MRI machine create some discrepancies between normal and MRI modes of articulation that cannot be evaluated with precision. Even if they are more flexible than concatenation of uniform tubes they require prior adaptation before being used for any speaker.

The number of parameters of an articulatory model generally ranges from 6 to 9 and the solution space cannot be explored during the inversion. Inversion methods therefore exploit some explicit or implicit table lookup method to recover at each time frame the set of articulatory parameters. Explicit table lookup methods required efficient sampling and representation methods [6] to limit the table size. Implicit table lookup methods often use neural networks [9, 1] but cannot guarantee that a uniform acoustic resolution is achieved.

Once inverse solutions have been recovered at each time frame of the speech signal articulatory trajectories are then built from these local solutions by some optimal path search algorithm, generally dynamic programming [3]. Other methods exploit regularization techniques or physical constraints [4] to obtain smooth trajectories.

The evaluation of an acoustic-to-articulatory inversion procedure comprises two aspects. The first is the acoustical faithfulness and ensures that inverted data are able to reproduce a speech signal as close as possible to the original. The closeness is generally evaluated by measuring the distance between original and synthetic formant frequencies.

The second aspect is that of the articulatory faithfulness. Unlike domains where data can be acquired easily (automatic speech recognition for instance) acquisition here requires medical imaging techniques which are often expensive (MRI, X-ray, electromagnetic articulography), hazardous (X-ray), perturb articulation (noise produced by medical machine, MRI especially), not fast enough to capture continuous speech (MRI), not precise enough (electromagnetic articulography). This explains that very few data are available all the more because some are required to build or adapt the articulatory model.

Current inversion techniques mostly concern vowels and sequences of vowels for one speaker. This domains thus necessitates substantial efforts to provide a general purpose inversion framework.

Given these results perspectives concern the incorporation of additional constraints in order to reduce the under-determination of the inverse problem. These constraints could be static and provided by phonetics to penalize unrealistic vocal tract shapes given a formant 3-tuple. But they could be dynamic and provided by computer vision techniques used to track visible articulators, i.e. lips and lower jaw. The additional knowledge of visible articulators gives rise to two or three articulatory parameters (jaw position, lip aperture and protrusion) and therefore considerably reduces the under-determination of the problem. This corresponds to a multimodal audio-visual approach of the inverse problem.

The second perspective is the use of standard spectral parameters as input data and the development of a general inversion method for all the classes of speech sounds. This is a hard problem and requires to be able to derive the acoustical behavior of the articulatory model from a standard model whatever the geometrical characteristics of an arbitrary speaker. Furthermore, this implicitly requires the development of a general articulatory model that works for consonants (voiced or unvoiced) as well as for vowels.

Bibliography

[1] G. Bailly, C. Abry, R. Laboissièrre, P. Perrier, and J.-L. Schwartz. Inversion and speech recognition. In J. Vandewalle, R. Boite, M. Mooner, and A. Osterlinck, editors, *Signal processing VI: Theories and Applications*, volume 1, pages 159–164, Brussels, Belgium, August 1992. Elsevier.

[2] O. Engwall. Modelling of the vocal tract in three dimensions. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, pages 113–116, Budapest, September, 1999.

[3] S. K. Gupta and J. Schroeter. Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis. *Journal of Acoustical Society of America*, 94(5):2517–2530, Nov 1993.

[4] Schoentgen J. and S Ciocea. Kinematic formant-to-area mapping. *Speech Communication*, 21:227–244, 1997.

[5] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.

[6] Slim Ouni and Yves Laprie. Exploring the Null Space of the Acoustic-to-Articulatory Inversion Using a Hypercube Codebook. In *Eurospeech, Aalborg, Danemark*, volume 1, pages 277–280, September 2001.

[7] R.C. Rose, J. Schroeter, and M.M. Sondhi. An investigation of the potential role of speech production models in automatic speech recognition. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 575–578, Yokohama, Japan, September, 1994.

[8] J. Schroeter and M. M. Sondhi. Speech coding based on physiological models of speech production. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 231–267. Dekker, New York, 1992.

[9] K. Shirai and T. Kobayashi. Estimating articulatory motion from speech wave. *Speech Communication*, 5(2):159–170, 1986.

Chapter 6

Speech segmentation/classification

C. Kotropoulos
AUTH, Greece

6.1 Introduction

Automatic segmentation (or endpoint detection) replaces efficiently human at the monotonous job of separating target sound units within audio files. The target might be speech, noise, silence, music [1], animal vocalization, vowels, and consonants. The audio segmentation is followed by automatic classification, which is used to categorize audio in the aforementioned categories. Audio segmentation and classification are crucial in various types of applications such as telecommunications, Automatic Speech Recognition (ASR), speech enhancement, and database construction.
The Voice Activity Detection (VAD) is a tool which is used broadly for separating human voice from a variety of noises and silence segments. In telecommunications, VAD is saving bandwidth because it prevents non-speech sounds to be transmitted. Some of the VAD algorithms are settled as standards. A widely used standard is the VAD of G.729 Annex B which is considered to be the state-of-the-art voice activity detector.

6.2 Procedure

The algorithms for audio segmentation and classification break a sound file in small pieces and give a prediction about what that segment might be. The small pieces of audio are called frames and usually have a duration of 10 to 350 msec. From the frame are derived some descriptive properties which are called features. Common features are the number of zero-crossings (passes from zero point), energy (logarithmic or Euclidean), the spectrum [6], the cepstrum, the Linear Prediction Coefficients, duration [8], and the formants. The G.729B VAD reference standard uses the full band energy, the low band energy, the zero-crossing rate and a spectral measure.
From the features is made a decision about the content of the frame. In early developments the decision was based only to the current frame. As the technology became faster, the decision algorithms evolved a memory which relates the prediction of the current frame with the previous frames. These algorithms are called adaptive. Some of these techniques, update a simple threshold [5], or adapt continously the parameters in the Hidden Markov Model (HMMs) [9]. An importand statistical theoretic framework is developed by J. Sohn [7].
Briefly, some of evolved adaptive algorithms are:

- Hidden Markov Models (HMMs): The HMMs retain time information about the probability distribution of the frame features. The probability distribution functions of the features are modelled usually as mixtures Gauss distributions. The decision is based on the simple Bayes rule [4].

- Spectrum transformation based on the properties of the human cochlea [3].
- Non-linear transformation of the spectrum [2].

All the aforementioned techniques are tested on databases. A common reference database is TIMIT speech database [4]. In the speech files is added noise in order to measure the robustness of the classification algorithm under various Signal to Noise Ratios (SNRs). Common SNRs are 25, 15, and 5 Decibels (dB). Most of the new VAD algorithms are compared with the G.729B [7], [10] which achieves the scores reported in Table 6.1.
Speech segmentation is considered as a front-end easy to establish procedure. The classification of the speech is more difficult than speech segmentation but the recently developed techniques are very efficient.

[b]

Enviroments		$P_d(\%)$			$P_f(\%)$
Noise	SNR	Voiced	Unvoiced	Speech	Noise
White noise	5 dB	75.62	15.18	64.14	1.01
	15 dB	93.42	50.83	85.33	1.63
	25 dB	99.07	85.15	96.43	3.56

Table 6.1: P_d, P_f: are speech detection and false alarm probabilities for the G.729B VAD.

Bibliography

[1] J. Ajmera, I. McCowan, and H. Bourland, “Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication*, no. 40, pp. 351-363, 2003.

[2] P. N. Garner, T. Fukadam, and Y. Komori, “A differential spectral voice activity detector," in *Proc. Inter. Conf. on Acoustics, Speech and Signal Proc. (ICASSP) 2004*, vol. 1, pp. 597-600, May 2004.

[3] N. Mesgarani, S. Shamma, and M. Slaney, “Speech discrimination based on multiscale spectro-temporal modulations," in *Proc. Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP) 2004*, vol. 1, pp. 601-604, May 2004.

[4] B. L. Pellom and J. H. L. Hansen, “Automatic segmentation of speech recorded in uknown noisy channel characteristics," *Speech Communication*, vol. 25, pp. 97-116, 1998.

[5] P. Sovka and P. Pollak, “The study of speech/pause detectors for speech enhancement methods," in *Proc. Eurospeech 95*, pp.1-4, Madrid, September 1995.

[6] M. Sharma and R. Mammone, “Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge," in *Proc. of IEEE Int. Conf. on Spoken Language Proc. (ICSLP) 1996*, Philadephia, PA, October 1996.

[7] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection," in *IEEE Signal Proc. Letters*, vol. 6, no. 1, January 1999.

[8] D. Wang, L. Lu, and H. J. Zhang, “Speech segmentation without speech recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP) 2003*, vol. I, pp. 468-471, Hong Kong, April 4-10, 2003.

[9] J. Ziang, W. Ward, and B. Pellom, “Phone based voice activity detection using online bayesian adaptation with conjugate normal distributions," on *Proc. Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP) 2002*, Orlando, Florida.

[10] ITU-T Rec. G.729, Annex B, “A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70".

[11] The Segmentation Task: Find the Story Boundaries, http://www.nist.gov/speech/tests/tdt/tdt99/presentations/NIST_segmentation/

[12] Natural Language Processing Group, http://torch.cs.dal.ca/ nlp/

[13] The Center for Spoken Language Research, http://cslr.colorado.edu/

[14] Speex, a free codec for free speech, http://www.speex.org/

[15] The International Engineering Consortium (IEC), topic on Voice Activity Detection, http://www.iec.org/online/tutorials/vocable/topic06.html

[16] The Infant Speech Segmentation Project, http://www-gse.berkeley.edu/research/completed/InfantSpeech.html

[17] RWCP Sound Scene Database in Real Acoustical Environments, Voice Activity Detection in Noisy Environments, http://tosa.mri.co.jp/sounddb/nospeech/research/indexe.htm

[18] Language Science Research Group, Washigton Univ., online demos, http://lsrg.cs.wustl.edu/

[19] UCL Psychology Speech Group, speech segmentation issues, http://www.speech.psychol.ucl.ac.uk/segmentation.html

Chapter 7

Speech / Music Segmentation

D. Burshtein, A. Yeredor
TAU-Speech, Israel

7.1 State of Art

Discrimination between speech and music is an important feature in many applications. With the increasing amounts of digital audio information and multimedia databases for storing it, there is a need for an automatic classifier which can allow fast indexing and retrieving of the stored data. In automatic speech recognition (ASR) applications, such a classifier can reject non speech segments. Another example for the use of music speech discriminator is as a preprocessing element prior to the application of low bit rate audio coder that relies on the characteristics of the audio signal to achieve better compression ratios.

A variety of systems for audio segmentation have been proposed in the past [1, 2, 3, 4, 5, 6, 7], relying on different types of features extracted from the audio signal. These features can be divided into three major categories.

The first category is based on the time domain and includes features such as zero crossing rate (ZCR) [10] and amplitude modulation. In [6] a speech / music classifier based on these two features is presented. First, the data is divided into segments according to significant changes in the signal's root-mean-square (RMS) distribution. Music and speech have different RMS distributions so these changes imply on probable transition in audio type. Each segment is then classified by a set of tests using features computed from the basic features RMS and ZCR. A success rate of about 95% is reported.

The second category is based on the frequency domain and includes features which are often used for ASR. Very popular features are the MEL frequency cepstral coefficients (MFCC), which were shown to be also good for music modeling [7]. Better results are obtained by using also the delta coefficients (delta MFCC) [2]. In [2] a comparison of four features was made for speech and music discrimination: Amplitude, zero crossing, pitch and MFCC. The results show that MFCC and delta MFCC gave the best performance.

The third category of features combines both the frequency and the time domains. These features usually focus on the changes in time of the spectrum like 4 Hz modulation energy [3] and spectral flux [1].

Most of the classifiers use one of the following frameworks: Gaussian mixture model (GMM) [8], K-nearest neighbors (KNN) and the hidden Markov model (HMM) [9]. A comparison of 13 different features was made in [1]. GMM, KNN and their simpler versions (k-d trees and Gaussian MAP) were used. It is shown that a small subset of features get the best results with a success rate of 94.2% for 20 ms segments and 98.6% for 2.4 sec segments. The comparison also shows no significant differences between the used models.

The work in [1] used a database containing digital audio samples of various radio stations. Other studies used commercial databases like the Waxholm Database (see [3]) and the King database (see [8]) for speech and various music types from music CDs.

7.2 Our Research: Planned Work

In our work we will conduct a comparative study between the various features and will examine combinations of different types of features by using short term (per frame) features and long term (per set of frames) features. Short term features will be evaluated by a GMM classifying each frame as music / speech according to the likelihood ratio. Different feature types will have separate models and the overall decision will be based on a combination of these models. The long term features will be used in the segmentation process as a smoothing procedure on the short term classification.

Bibliography

[1] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," in Proc. ICASSP'97, Munich, Vol. II, pp. 1331–1334, 1997.

[2] M. J. Carey, E. S. Parris and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *Proc. ICASSP99*, pp. 149–152, Phoenix, 1999.

[3] S. Karneback, "Discrimination between speech and music based on a low frequency modulation feature," in Proc. Eurospeech, 2001 (available at `http://www.speech.kth.se/~stefank`).

[4] J. Pinquier, J. L. Rouas and R. Andre-Obrecht, "Robust speech / music classification in audio documents," in *Proc. ICSLP*, pp. 2005-2008, 2002.

[5] W. Q. Wang, W. Gao and D. W. Ying, "A Fast and Robust Speech/Music Discrimination Approach," *The Fourth International Conference on Information, Communications & Signal Processing and Fourth Pacific-Rim Conference on Multimedia (ICICS-PCM 2003)*, Singapore, December 15–18, 2003 (available at `http://www.jdl.ac.cn/doc/2003`)

[6] C. Panagiotakis and G. Tziritas, "A Speech/Music Discriminator Based on RMS and Zero-Crossings," *IEEE Transactions on Multimedia*, 2003 (available at `http://www.csd.uch.gr`).

[7] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," in *Proc. of the International Symposium on Music Information Retrieval*, 2000 (available at `http://ciir.cs.umass.edu/music2000/papers`).

[8] D. Reynolds and R. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

[9] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[10] B. Kedem, "Spectral Analysis and Discrimination by Zero-Crossings," *Proceedings of the IEEE*, vol. 74 ,pp. 1475–1493, 1986.

Chapter 8

Speaker indexing in large audio archives

D. Burshtein, A. Yeredor
TAU-Speech, Israel

8.1 State of Art

Indexing large audio archives has emerged recently [2, 4] as an important research topic as large audio archives now exist. The goal of speaker indexing is to divide the speaker recognition process into two stages. The first stage is a pre-processing phase which is usually done on-line as audio is recorded into the archive. In this stage there is no knowledge about the target speakers. The goal of the pre-processing stage is to do all possible pre-calculations in order to make the search as efficient as possible when a query is presented. The second stage is activated when a target speaker query is presented. In this stage the pre-calculations of the first stage are used. Contrary to speaker indexing, state-of-the-art speaker recognition algorithms usually perform most of the computation after an audio sample of a target speaker is presented and a model is estimated to the target speaker.

The motivation for speaker indexing is that classic speaker recognition algorithms are not efficient enough for scanning large audio archives. A speaker indexing algorithm must be time efficient, it must require only little additional storage space and still be accurate. Another appealing attribute may be to enable searching for a speaker without accessing the audio itself, either because the audio is stored in a slow storage device or because it has been already erased.

Possible applications for using speaker indexing are audio search engines and speech mining for CRM.

8.1.1 Literature survey on Speaker recognition

For almost a decade, Gaussian Mixture Models [6] (GMMs) are used by most state-of-the-art speaker recognition algorithms. The basic idea of the GMM based speaker recognition algorithm is to model a sequence of acoustic vectors (frames) extracted from an utterance by assuming frame independence and modelling the likelihood of an acoustic vector given a speaker by a mixture of normal distributions. Usually most of the research done in speaker recognition attempts to improve the accuracy of recognition. However, there are also some important computational issues related to speaker recognition. In [5] several ways to improve time complexity of the GMM algorithm are explored. In [7] the issue of compressing the size of a GMM was addressed.

8.1.2 Literature Survey on Speaker indexing

In [8] it is suggested to perform speaker indexing by projecting each utterance into a speaker space defined by anchor models which are a set of non-target speaker models(GMMs). Each utterance is represented by a vector of distances between the utterance and each anchor model. This representation

is calculated in the pre-processing phase. In the query phase, the target speaker data is projected to the same speaker space and the speaker space representation of each utterance in the archive is compared to the target speaker vector, using a distance measure such as Euclidean distance. The disadvantage of this approach is that it is suboptimal. Indeed, the EER reported in [8] is almost tripled when using anchor models instead of conventional GMM scoring. This disadvantage was handled by cascading the anchor model indexing system and the GMM recognition system thus first filtering efficiently most of the archive and then rescoring in order to improve accuracy. Nevertheless, the cascaded system failed to obtain accurate performance for speaker mis-detection probability lower than 50%. Another drawback of the approach is that sometimes the archive is not accessible for the search system either because it is too expensive to access the audio archive, or because the audio itself was deleted from the archive because of lack of available storage resources (the information that a certain speaker was speaking in a certain utterance may be beneficial even if the audio no longer exists, for example for law enforcement systems). Therefore, it may be important to be able to achieve accurate search with low time and memory complexity using only an index file without the raw audio.

8.2 Our Research

8.2.1 Speaker indexing using test utterance Gaussian mixture modelling

In [1] we present a novel approach for speaker indexing. The idea is to compute the likelihood of a test utterance given a GMM indirectly. Instead of scoring each frame of the test utterance given the GMM, we compute the likelihood in a way that distributes the computation into two stages. In the first stage (which is independent of the target speaker) a GMM is fitted to the test utterance. In the second stage the likelihood of the test utterance given a target speaker GMM is calculated using the GMM representation of the test utterance instead of the feature vectors. We present an efficient algorithm to approximate the likelihood of the test utterance given a target speaker GMM by using the GMM representation of the test utterance instead of the feature vectors. Testing our algorithm on the SPIDRE corpus [3] shows that the approximation does not reduce accuracy. Using this method reduced the time needed for searching for a speaker by a factor of 125 and needs an overhead of 7% in storage.

8.2.2 Planned Work

Our ongoing research focuses on 2 aspects. The first one is to reduce the overhead storage required, and the second one is to further accelerate the search algorithm (stage 2). The reduction in storage required can be achieved by quantization and compression of the GMMs fitted to each utterance in the archive. The acceleration in search time can be achieved by further improving our algorithm and by treating the test utterances as vectors in a high-dimensional GMM space and searching efficiently in that space using techniques developed in the discipline of searching and indexing in high-dimensional spaces.

Bibliography

[1] H. Aronowitz, D. Burshtein, and A. Amir. Speaker indexing in audio archives using test utterance Gaussian mixture modeling. To appear in ICSLP '04.

[2] I. M. Chagolleau and N. P. Valles. Audio indexing: What has been accomplished and the road ahead. In *JCIS*, pages 911–914, 2002.

[3] Linguistic Data Consortium. *SPIDRE documentation file*. `http://www.ldc.upenn.edu/Catalog/readme_files/spidre.readme.html`.

[4] J. Foote. An overview of audio information retrieval. *ACM Multimedia Systems*, (7):2–10, 1999.

[5] J. McLaughlin, D. A. Reynolds, and T. Gleason. A study of computation speed-ups of the gmm-ubm speaker recognition system. In *Proc. Eurospeech '99*, pages 1215–1218, Rhodes, Greece, 1999.

[6] D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17:91–108, August 1995.

[7] D. A. Reynolds. Model compression for gmm based speaker recognition systems. In *Proc. Eurospeech '03*, pages 2005–2008, 2003.

[8] D. E. Sturim, D. A. Reynolds, E. Singer, and J. P. Campbell. Speaker indexing in large audio databases using anchor models. In *Proc. ICASSP '01*, pages 429–432, 2001.

Chapter 9

Audio Indexing

A. Rauber
VUT, Austria

9.1 Introduction

A significant amount of research has been conducted in the area of content-based music retrieval, focusing on different characteristics and information needs with respect to access to audio collections. There is a variety of approaches to music indexing and there are many related disciplines involved. Because of such wide varieties, it is difficult to cite all the relevant work. Current approaches to music indexing can broadly be classified into data-based and content-based approaches. For the aims of scientific research on multimedia indexing, content-based approaches are more interesting, nevertheless the use of auxiliary textual data structures, or metadata, can frequently be observed in approaches to non-textual, e.g. image or video document indexing. Indeed, textual index terms are often manually assigned to multimedia documents to allow users retrieving documents through textual descriptions.

Methods have been developed to search for pieces of music with a particular melody. In query-by-humming approaches (QBH) users may formulate a query by humming a melody, which is then usually transformed into a symbolic melody representation. This is matched against a database of scores given, for example, in MIDI format. Other approaches focus on the sound characteristics of music, trying to groups pieces of music by musical genre or sound similarity. Yet another approach with respect to audio retrieval and access focuses on speech, rather than music sounds. We will cover these different fields in more detail below, starting with QBH systems, followed by a description of the major contributions in the field of genre-oriented access, rounding up with a short summary on speech-related research for access (rather than the more dominant speech recognition domain).

9.1.1 Symbolic Music Indexing and Retrieval

Sometimes music is available in the MIDI format [24] and for a limited subset of music automatic transcription to the MIDI format is possible (e.g. [25, 21]). Music archives in MIDI format contain information on the exact melody for each song. Hawley presented a system [15], where the user enters a melody on a keyboard and tunes whose beginnings exactly match the input are retrieved. Ghias et al. [12] presented a system where the user hummes a query, which is reduced to relative in-formation such as if a note is higher, lower or about the same as the previous, and retrieves songs which have similar melodies.

One of the best known representatives of such systems is the New Zealand Musical Digital Library [22, 1]. Additionally to the melody of a piece of music information on its style can be abstracted from the MIDI data. Dannenberg et al. [4] described a system, which classified solo improvised trumpet performances into one of the four styles: lyrical, frantic, syncopated, or pointillistic.

Another approach, reported in [16], applies pattern matching techniques to documents and queries in GUIDO format, exploiting the advantages of this notation in structuring information. Approximate

string matching has been used also by [14]. Markov chains have been proposed in [2] to model a set of themes that has been extracted from music documents, while an extension to hidden Markov models has been presented in [31] as a tool to model possible errors in sung queries.

In [6] melodies were indexed through the use of N-grams, each N-gram being a sequence of N pitch intervals. Experimental results on a collection of folk songs were presented, testing the effects of system parameters such as N-gram length, showing good results in terms of retrieval effectiveness, though the approach seemed not be robust to decreases in query length. Another approach to document indexing has been presented in [23], where indexing has been carried out by automatically highlighting music lexical units, or musical phrases. Differently than the previous approach, the length of indexes was not fixed but depended on the musical context. That is musical phrases were computed exploiting knowledge on music perception, in order to highlight only phrases that had a musical meaning. Phrases could undergo a number of different normalization, from the complete information of pitch intervals and duration to the simple melodic profile.

9.1.2 Subsymbolic Music Indexing and Retrieval

However, most music is available in MP3 or other similar formats rather than in MIDI. Thus content-based systems which directly analyze the raw music data (acoustical signals) have been developed. The models of these systems for music processing are usually sub-symbolic [17] since they operate directly on acoustical signals and do not involve an explicit symbolization step as, for example, in [34], where the musical structure is analyzed on a very abstract level. An overview of systems analyzing audio databases was presented by Foote [10]. However, Foote focuses particularly on systems for retrieval of speech or partly-speech audio data. Several studies and overviews related to content-based audio signal classification are available (e.g. [19]), however, they do not treat content-based music classification in detail. One of the few exceptions to this is presented in [18], where hummed queries are posed against an MP3 archive for melody-based retrieval.

Up till now only few approaches in the area of content-based music analysis have utilized the framework of psychoacoustics. Psychoacoustics deals with the relationship of physical sounds and the human brain's interpretation of them, cf. [36]. One of the first exceptions was [7], using psychoacoustic models to describe the similarity of instrumental sounds. The authors used a collection of instrument sounds, which were organized using a Self-Organizing Map in a similar way as presented in this paper. For each instrument a 300 milliseconds sound was analyzed, extracting steady state sounds with a duration of 6 milliseconds. These steady state sounds can be regarded as the smallest possible building blocks of music.

Wold et al. [35] presented a system which analyzes sounds based on their pitch, loudness, brightness, and bandwidth over time and tracked the mean, variance, and autocorrelation functions of these properties. They analyze sounds such as speech and individual musical notes, but do not focus on whole music collections. Other approaches (e.g. [11]) are based on methods developed in the digital speech processing community using Mel Frequency Cepstral Coefficients (MFCCs). MFCCs are motivated by perceptual and computational considerations, for example, instead of calculating the exact loudness sensation only decibel values are used. Furthermore the techniques appropriate to process speech data are not necessarily the best for processing music. For example, the MFCCs ignore some of the dynamic aspects of music. Recently Scheirer [30] presented a model of human perceptual behavior and briefly discussed how his model can be applied to classifying music into genre categories and performing music similarity-matching. However, he has not applied his model to large scale music collections. The collection he uses consisted of 75 songs from each of which he selected two 5-second sequences.

A more practical approach to this task was presented in [33] where music given as raw audio is classified into genres based on musical surface and rhythm features. Another system supporting browsing audio streams is the Sonic Browser [9]. In [8] this line of research is continued evaluating the browsing of everyday sounds. The investigation is directed at comparing browsing single versus multiple stream

audio.

A completely different approach is taken by the SOMeJB system, i.e. the SOM-enhanced Jukebox [28]. The goal is to automatically create an organization of music archives following their perceived sound similarity. More specifically, characteristics of frequency spectra are extracted and transformed according to psychoacoustic models. The resulting psychoacoustic Rhythm Patterns are further organized using the Growing Hierarchical Self-Organizing Map (GHSOM), an unsupervised neural networkbased on the Self-Organizing Map. On top of this advanced visualizations including Islands of Music (IoM) and Weather Charts [27] offer an interface for interactive exploration of large music repositories.

9.2 URLs

[3, 13, 26, 32, 29, 20]

9.3 Overview Article

[5]

Bibliography

[1] D. Bainbridge, C.G. Nevill-Manning, H. Witten, L.A. Smith, and R.J. McNab. Towards a digital library of popular music. In E.A. Fox and N. Rowe, editors, *Proceedings of the ACM Conference on Digital Libraries (ACMDL'99)*, pages 161–169, Berkeley, CA, August 11-14 1999. ACM. `http://www.acm.org/dl`.

[2] W.P. Birmingham, R.B. Dannenberg, G.H. Wakefield, M. Bartsch, D. Bykowski, D. Mazzoni, C. Meek, M. Mellody, and W. Rand. MUSART: Music retrieval via aural queries. In *Proceedings of the 2nd Annual Symposium on Music Information Retrieval (ISMIR 2001)*, Bloomington, ID, October 15-17 2001. `http://ismir2001.indiana.edu/papers.html`.

[3] Cantate. computer access to notation and text in music libraries. Website, 2004. `http://projects/fnb/nl/cantate`.

[4] R.B. Dannenberg, B. Thom, and D. Watson. A machine learning approach to musical style recognition. In *Proceedings of the International Computer Music Conference (ICMC 97)*, pages 344–347, Thessaloniki, Greece, September 25-30 1997. `http://www-2.cs.cmu.edu/~rbd/bib-styleclass.html#icmc97`.

[5] J.S. Downie. *Annual Review of Information Science and Technology*, volume 37, chapter Music information retrieval, pages 295–340. Information Today, Medford, NJ, 2003. `http://music-ir.org/downie_mir_arist37.pdf`.

[6] S. Downie and M. Nelson. Evaluation of a simple and effective music information retrieval method. In *Proceedings of the 23nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR00)*, pages 73–80, Athens, GR, 2000. ACM. `http://www.acm.org/dl`.

[7] B. Feiten and S. Günzel. Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal*, 18(3):53–65, 1994.

[8] M. Fernström and E. Brazil. An auditory tool for multimedia asset management. In *Proceedings of the International Conference on Auditory Display (ICAD 2001)*, Espoo, Finland, July 29 - August 1 2001.

[9] M. Fernström and C. McNamara. After direct manipulation - direct sonification. In *Proceedings of the International Conference on Auditory Display (ICAD 98)*, Glasgow, UK, 1998.

[10] J. Foote. An overview of audio information retrieval. *Multimedia Systems*, 7(1):2–10, 1999. `http://www.fxpal.com/people/foote/papers/index.htm`.

[11] J.T. Foote. Content-based retrieval of music and audio. In C.-C.J. Kuo, editor, *Proceedings of SPIE Multimedia Storage and Archiving Systems II*, volume 3229, pages 138–147, 1997. `http://www.fxpal.xerox.com/people/foote/papers/spie97-abs.html`.

[12] A. Ghias, J. Logan, D. Chamberlin, and B.C. Smith. Query by humming: Musical information retrieval in an audio database. In *Proceedings of the Third ACM International Conference on Multimedia*, pages 231–236, San Francisco, CA, November 5 - 9 1995. ACM. http://www.acm.org/dl.

[13] Harmonica. accompagnying action on music information in libraries. Website, 2004. http://projects/fnb/nl/harmonica.

[14] G. Haus and E. Pollastri. A multimodal framework for music inputs. In *Proceedings of thze ACM Multimedia 2000 Conference*, pages 282–284, Plymouth, USA, 2000. ACM.

[15] M. Hawley. The personal orchestra. *Computing Systems*, 3(2):289–329, 1990.

[16] H.H. Hoos, K. Renz, and M. Görg. GUIDO/MIR - an experimental music information retrieval system based on GUIDO music notation. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2001)*, pages 41–50, Bloomington, USA, 2001.

[17] M Leman. Symbolic and subsymbolic information propcessing in models of musical communication and cognition. *Interface*, 18:141–160, 1989.

[18] C.-C. Liu and P.-J. Tsai. Content-based retrieval of mp3 music objects. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM 2001)*, pages 506 – 511, Atlanta, Georgia, 2001. ACM. http://www.acm.org/dl.

[19] M. Liu and C. Wan. A study of content-based classification and retrieval of audio database. In *Proceedings of the 5th International Database Engineering and Applications Symposium (IDEAS 2001)*, Grenoble, France, 2001. IEEE.

[20] Marsyas: A software framework for research in computer audition. Website. http://www.cs.princeton.edu/~gtzan/wmarsyas.html.

[21] R.J. McNab, Smith L.A., and I.H. Witten. Signal processing for melody transcription. In *Proceedings of the 19th Australasian Computer Science Conferences*, pages 301–307, Melbourne, Australia, 1996.

[22] R.J. McNab, L.A. Smith, J.H. Witten, C.L. Henderson, and S.J. Cunningham. Towards the digital music library: Tune retrieval from acoustic input. In *Proceedings of the 1st ACM International Conference on Digital Libraries*, pages 11–18, Bethesda, MD, USA, March 20 - 23 1996. ACM. http://www.acm.org/dl.

[23] M. Melucci and N. Orio. Musical information retrieval using melodic surface. In E.A. Fox and N. Rowe, editors, *Proceedings of the ACM Conference on Digital Libraries (ACMDL'99)*, pages 152–160, Berkeley, CA, August 11 - 14 1999. ACM. http://www.acm.org/dl.

[24] MIDI Manufacturers Association (MMA). MIDI 1.0 Specification, V 96.1. online, March 1996. http://www.midi.org.

[25] J.A. Moorer. On the transcription of musical sound by computer. *Computer Music Journal*, 1(4):32–38, 1977.

[26] Musica. the international database of choral repertoire. Website, 2004. http://www.musicanet.org.

[27] E. Pampalk, A. Rauber, and D. Merkl. Content-based organization and visualization of music archives. In *Proceedings of ACM Multimedia 2002*, pages 570–579, Juan-les-Pins, France, December 1-6 2002. ACM. http://www.ifs.tuwien.ac.at/ifs/research/publications.html.

[28] A. Rauber, E. Pampalk, and D. Merkl. The SOM-enhanced JukeBox: Organization and visualization of music collections based on perceptual models. *Journal of New Music Research*, 32(2):193–210, June 2003. http://www.extenza-eps.com/extenza/loadHTML?objectIDValue=16745&type=ab%stract.

[29] RISM. Repertoire international des sources musicales. Website, 2004. http://rism.stub.uni-frankfurt.de.

[30] E.D. Scheirer. *Music-Listening Systems*. PhD thesis, MIT Media Laboratory, 2000. http://web.media.mit.edu/~eds/thesis/.

[31] J. Shifrin, B. Pardo, C. Meek, and W. Birmingham. HMM-based musical query retrieval. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2002)*, pages 295–300, Portland, OR, 2002.

[32] SOMeJB: The SOM-enhanced jukebox. Website. http://www.ifs.tuwien.ac.at/~andi/somejb.

[33] G. Tzanetakis, G. Essl, and P.R. Cook. Automatic musical genre classification of audio signals. In *Proceedings International Symposium on Music Information Retrieval (ISMIR)*, pages 205–210, Bloomington, Indiana, October 15-17 2001. http://ismir2001.indiana.edu/papers.html.

[34] G. Widmer. Using AI and machine learning to study expressive music performance: Project survey and first report. *AI Communications*, 14(3), 2001.

[35] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification search and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, Fall 1996.

[36] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models*, volume 22 of *Series of Information Sciences*. Springer, Berlin, 2 edition, 1999.

Chapter 10

Event Detection, Segmentation and Classification for Audio Streams

D. Dimitrios, P. Vassilis and P. Maragos
ICCS-NTUA, Greece

10.1 Introduction

Scientific and technological progress nowadays facilitate the continuous accumulation of data. This fact yields unprecedented opportunities for the transition of our society to the prospect of a knowledge-driven society. Semantic manipulation of the deluge of information that is being accrued is a basic prerequisite towards this direction. Diversity and multimodality of this information pose additional challenges.

Within this framework, tackling automatic content analysis of audio data is of major importance. Audio cues, either alone or integrated with information extracted from other modalities, may contribute significantly to the overall semantic interpretation of data.

Event detection in audio streams is an aspect of the aforementioned analysis. The concept of "*event*" corresponds to a noteworthy happening and is application dependent. For example, applause, chanting, laughter or alterations in the speech rate of the sport-caster may be regarded as events in a sports-video. Speaker changes, changes between various speech quality levels, between speech and silence, or speech and music are common events in the case of broadcast news.

Event detection in audio streams aims at delineating audio as a sequence of homogeneous parts each one identified as a member of a predefined audio class. *Determination of such audio classes* (e.g. speech, music, silence, laughter, noisy speech etc.) is the first step in the design of an event detection algorithm. Obviously, these classes are inferred from the specificity of the application. The *selection and extraction of appropriate audio-features* ensues and is probably the most significant phase. It is understood that the correct choice of these features favors the successful completion of the following phases. Proper exploitation of the extracted attributes leads to the *segmentation* of the audio stream and *classification* of the resulting parts.

In this report we will attempt to discuss the state of the art in all these aspects of event detection, namely selection of audio features, segmentation and classification. Research in these areas has been quite intense and indicates that the problem is an arduous one.

A relevant overview article has been written by Foote, [10].

10.2 Audio Feature Selection

During the short life of research in the area of general audio segmentation various types of features have been proposed. These span categories from the standard features used in Automatic Speech Recognition (ASR) (e.g. they can represent short-time spectral/envelope information or frequency content via zero-

crossings), to various specially designed features that try to account for specific characteristics of each audio type, or that can discriminate among them. Another case of specially designed features are the ones that discriminate between two classes of audio data, e.g. common discrimination tasks include speech vs music or speech vs silence.

The standard features used in ASR (Mel Frequency Cepstral Coefficients, Perceptual Linear Prediction Cepstral Coefficients) have been adopted in many approaches. Most of them afterwards use model-based (e.g. posterior probabilities) classification methods [32, 20, 5, 9, 29] and/or are designated to be part of larger ASR systems. Other standard popular features used in [34], [23], [14], [22], [19] include short-time *zero-crossings rate* (ZCR), related to the mean frequency of a segment, and *s*hort-time energy (STE) that provides a convenient representation of the amplitude. Variations of these are *high zero-crossing rate ratio* defined as the ratio of the number of frames whose ZCR is above 1.5 times the average ZCR, and *low short-time energy ratio* which is proportional to the ratio of the number of frames whose STE is less than 0.5 time of average STE. *S*ilence crossing rate is the number of times that the energy falls below some silence level criterion. Another proposed feature is the *Spectrum Flux* defined as the average variation value of spectrum between two adjacent frames. Even the root mean square amplitude along with the ZCR has been used as feature in [23] resulting in quite high classification rates.

Among the features that have been proposed recently especially for the specific problem of audio segmentation and classification , the following should be mentioned: *average probability dynamism*, *mean per frame entropy* [30], [2], [5], *background-label energy ratio, phoneme distribution match* [30], *Teager energy* and related *modulation features* [8].

In more detail, the use of *average probability dynamism* as an audio-feature is based on the observation that the posterior phoneme probability estimates (according to an acoustic model) for speech segments change frequently whereas in non-speech segments they change much less frequently and more gradually. *Mean per-frame entropy* exhibits the suitability of an acoustic model for an audio segment. *Background-label energy ratio* compares the expected non-speech energy of a segment to the expected speech energy. *Phoneme distribution match* shows how close the phoneme probability distribution for a whole segment is to the corresponding distribution estimated from a training set of known speech segments. Finally, *Teager energy* and related *modulation features*, namely instantaneous amplitude and frequency, are based on the observation that speech resonance signals may be modelled as amplitude and frequency modulated signals.

There are few cases that the features selected might be characterized by unnecessary detail and they are not so general as they should be. As an example, in [25], various over-detailed quantities are proposed as features, like cepstrum resynthesis residual magnitude and spectral roll-off point.

Features that discriminate between speech and music should be mentioned explicitly, since they have been the object of specialized research. Most sounds generated by musical instruments as sources share the common characteristic of being harmonic. This means that they contain superimposed a fundamental frequency tone plus its integer multiples. On the other hand, speech is a mixed harmonic/non-harmonic sound with voice/unvoiced segments correspondingly [34]. Although fundamental frequency estimation (referred as pitch in speech) is a research field on its own, various features try to detect harmonic related properties.

10.3 Segmentation

Many different approaches have been proposed in the literature for the segmentation of an audio stream into homogeneous parts.

In *rule-based* approaches, segmentation is based on rules applied to the set of features that have been extracted from the stream. *Energy* is the most common feature used. Energy based approaches [28], [16], [14] have been widely used and are particularly easy to implement. Silence periods in the input signal are detected as low energy sections of the signal. It is assumed that segment boundaries exist in these periods if a number of additional constraints is satisfied such as minimum length of the silence period

. Others hypothesize segment boundaries when abrupt changes in the values of the features between subsequent moving frames are detected [33], [34].

In *metric-based* approaches segment boundaries are placed at local maxima or minima of a special distance calculated between neighboring sliding windows. One metric is the Kullback-Leibler divergence that was first used by Siegler et al [27] as an alternative to the Generalized Likelihood Ratio proposed in [12], in the case of speaker segmentation. The Bayesian Information Criterion (BIC) was applied as a metric by Chen and Gopalakrishnan [6] and exhibits improved stability and robustness at a high computational cost, however. Many have proposed variations in the application of the BIC in order to optimize efficiency, [35], [15]. The VQ distortion measure, on the other hand, proposed in [21], is an alternative reported to have improved results.

In *decoder-guided* approaches the speech recognition system is used for segmentation. The stream is first decoded and then the segments are cut between long silence intervals [17], [31]. However, silence is not directly related to acoustic changes in the stream so usually a second stage segmentation follows usually based on rules.

In many cases, segmentation is performed explicitly at pre-specified time intervals. A post processing step, after classification, follows in order to concatenate neighboring segments of the same class.

10.4 Classification

Equally important for event detection in audio streams is the classification of the various audio segments in predetermined classes. Classification may take place in subsequent stages. Definition of the classes depends on the application.

Rule-based methods follow a hierarchical heuristic scheme to achieve classification. Based on the properties of the various audio classes in the feature space, simple rules are devised and form a decision tree aiming at the proper classification of the audio segments [34], [14]. These methods usually lack robustness because they are threshold dependent, but no training phase is necessary and they can work in real-time.

In most of the *model-based* methods, segmentation and classification are performed at the same time. Models such as Gaussian Mixture Models and Hidden Markov Models are trained for each audio class and classification is achieved by Maximum Likelihood or Maximum a Posteriori selection over a sliding window [16], [4], [25], [24], [3], [1], [30]. These methods may yield quite good results but they cannot easily generalize, they do not work in real-time, since they usually involve a number of iterations, and data is needed for training.

Classical Pattern Analysis techniques cope with the classification issue as a case of pattern recognition. So, various well known methods of this area are applied, such as neural networks and Nearest Neighbor (NN) methods. Maleh et al [7] apply either a quadratic Gaussian classifier or an NN classifier. Shao et al [26] apply a multilayer perceptron combined with a genetic algorithm to achieve 16-class classification. Lu et al [19] apply an algorithm based on K-nearest neighbor classifier and Linear Spectral Pairs Vector Quantization to determine speech / non-speech segments. Foote [11] uses a tree structure quantizer for speech/music classification. More modern approaches have also been tested like Nearest Feature Line Method [18] which performs better than simple NN approaches, and Support Vector Machines [13], [20]. Results are quite satisfactory.

State of the art in the classification methods for event detection is not completely described by the preceding categorization. *Hybrid* approaches that combine the aforementioned ideas are equally significant. The classifier proposed by Lu et al [19] is such an example. In the first stage, a variation of classical pattern analysis algorithms is applied to discriminate between speech and non-speech segments, then a finer classification is achieved by a rule-based schema and finally speaker clustering is performed by model-based analysis.

10.5 Conclusions

In this report, we have presented briefly the state of the art in the area of event detection in audio streams. From the preceding discussion it becomes obvious that research in this field has been rather active in recent years. Although the achievements have been important, there is a number of key issues that still remain open. Generalization, robustness and real-time operation constitute challenging problems that ongoing research has to face. It seems that a widely accepted solution in the case of a general audio stream has not yet been proposed.

Further investigation of these areas is clearly warranted. Semantic interpretation of multimedia data will surely benefit from any possible improvements in the field of event detection in audio streams.

10.6 Related URLs

Related research projects:
`http://research.microsoft.com/users/llu/Audioprojects.aspx`

Audio search technologies:
`http://www.musclefish.com`

Audio mining technologies:
`http://www.nexidia.com/`
`http://www.bbn.com/speech/am.html`

Audio search using speech recognition:
`http://speechbot.research.compaq.com/`

Search the web for sounds:
`http://www.findsounds.com/index.html`

Musical audio mining:
`http://www.ipem.rug.ac.be/MAMI/`

Bibliography

[1] J. Ajmera, I. McCowan, and H. Bourlard. Robust HMM-based speech/music segmentation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-02)*, pages 1746–1749, Orlando, Florida, 2002.

[2] J. Ajmera, I. McCowan, and H. Bourlard. Speech/music segmentation using entropy and dynamism features in a HMM classification framework. *Speech Communication*, 40:351–363, 2003.

[3] M. Baillie and J.M. Jose. An audio-based sports video segmentation and event detection algorithm. In *Proc. 2nd IEEE Workshop on Event Mining 2004, Detection and Recognition of Events in video in association with IEEE Computer Vision and Pattern Recognition (CVPR2004)*, Washington DC, USA, July 2004.

[4] R. Bakis, S. Schen, P. Gopalakrishnan, R. Gopinath, S. Maes, and L. Polymenakos. Transcription of broadcast news - system robustness issues and adaptation techniques. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-97)*, volume 2, pages 711–714, 1997.

[5] A. Berenzweig and D. Ellis. Locating singing voice segments within music signals. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-01)*, 2001.

[6] S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Broadcast News Transcription and Understanding Workshop*, pages 127–132, Lansdowne, Virginia, February 1998.

[7] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal. Speech/music discrimination for multimedia applications. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-00)*, pages 2445–2448, June 2000.

[8] G. Evangelopoulos and P. Maragos. Multiband energy tracking and demodulation towards noisy speech endpoint detection. IEEE Trans. Acoust., Speech, Signal Processing, 2004. submitted.

[9] H. Ezzaidi and J. Rouat. Speech, music and songs discrimination in the context of handsets variability. In *Proc. International Conference on Speech and Language Processing, (ICSLP-02)*, September 2002.

[10] J. Foote. An overview of audio information retrieval. *Multimedia Systems, special issue on audio and multimedia*, 7(1):2–10, January 1999.

[11] J. T. Foote. Content-based retrieval of music and audio. In C.-C. J. Kuo et al., editor, *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, volume 3229, pages 138–147, 1997.

[12] H. Gish and M. Schmidt. Text-independent speaker identification. *IEEE Signal Processing Mag.*, pages 18–32, October 1994.

[13] G. D. Guo and S. Z. Li. Content-based audio classification and retrieval by Support Vector Machines. *IEEE Trans. Neural Networks*, 14(1):209–215, January 2003.

[14] H. Harb, L. Chen, and J.-Y. Auloge. Speech/music/silence and gender detection algorithm. In *Proceedings of the 7th International conference on Distributed Multimedia Systems DMS01*, pages 257–262, September 2001.

[15] R. Huang and J.H.L. Hansen. Advances in unsupervised audio segmentation for the broadcast news and NGSW corpora. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-04)*, 2004.

[16] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-00)*, 2000.

[17] F. Kubala, H. Jin, S. Matsoukas, L. Nguyen, R. Schwartz, and J. Makhoul. BBN byblos Hub-4 transcription system. In *Proc. of the 1997 DARPA Speech Recognition Workshop*, Chantilly, Virginia, 1997.

[18] S. Z. Li. Content-based audio classification and retrieval using the Nearest Feature Line Method. *IEEE Trans. Acoust., Speech, Signal Processing*, 8(5), September 2000.

[19] L. Lu, H.-J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *IEEE Trans. Acoust., Speech, Signal Processing*, (7):504–516, October 2002.

[20] P. J. Moreno and R. Rifkin. Using the Fisher Kernel Method for web audio classification. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-00)*, 2000.

[21] S. Nakagawa and K. Mori. Speaker change detection and speaker clustering using VQ distortion measure. *Systems and Computers in Japan*, 34(13):25–35, 2003.

[22] N. Nitanda, M. Haseyama, and H. Kitajima. An audio signal segmentation and classification using fuzzy c-means clustering. In *Proc. 2nd International Conference on Information Technology for Application (ICITA-2004)*, 2004.

[23] C. Panagiotakis and G. Tziritas. A speech/music discriminator based on RMS and zero-crossings. *IEEE Trans. Multimedia*, 2004 (accepted).

[24] J. Pinquier, J.-L. Rouas, and R. Andre'-Obrecht. Robust speech / music classification in audio documents. In *Proc. International Conference on Speech and Language Processing, (ICSLP-02)*, pages 2005–2008, Vol. 3, Denver, USA, 16-20 septembre 2002. Causal Productions Pty Ltd.

[25] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-97)*, 1997.

[26] X. Shao, C. Xu, and M. S. Kankanhalli. Applying neural network on content based audio classification. In *IEEE Pacific-Rim Conference On Multimedia (PCM-03), Singapore*, 2003.

[27] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proceedings of the Ninth Spoken Language Systems Technology Workshop*, 1997.

[28] H. Wactlar, A. Hauptmann, and M. Witbrock. Informedia: News-on-demand experiments in speech recognition. In *Proceedings of ARPA Speech Recognition Workshop*, Harriman, NY, February 1996. Arden House.

[29] Steven Wegmann, Francesco Scattone, Ira Carp, Larry Gillick, Robert Roth, and Jon Yamron. Dragon Systems' 1997 broadcast news transcription system. In *Proc. of the 1998 DARPA Broadcast News Workshop*, Landsowne, Virginia, 1998.

[30] G. Williams and D. Ellis. Speech/music discrimination based on posterior probability features. In *Proc. Eurospeech99*, Budapest, September 1999.

[31] P. C. Woodland. The development of the HTK broadcast news transcription system: An overview. *Speech Communication*, 37:47–67, May 2002.

[32] P.C. Woodland, T. Hain, G.L. Moore, T.R. Niesler, D. Povey, A. Tuerk, and E.W.D. Whittaker. The 1998 HTK broadcast news transcription system: Development and results. In *Proc. of the DARPA Broadcast News Workshop*, Herndon, Virginia, 1999.

[33] Wang Y., Liu Z., and Huang J. Multimedia content analysis using audio and visual information. *IEEE Signal Processing Mag.*, 17(6):12–36, November 2000. Invited Paper.

[34] T. Zhang and C.-C. J. Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Trans. Speech Audio Processing*, 9(4):441–457, May 2001.

[35] B. Zhou and J. H. L. Hansen. Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. In *Proc. International Conference on Speech and Language Processing, (ICSLP-00)*, pages 714–717, October 2000.